Optimierung von Messinstrumenten im Large-scale Assessment

Dissertation

zur Erlangung des akademischen Grades
Doctor rerum naturalium (Dr. rer. nat.)

im Fach Psychologie

eingereicht an der
Lebenswissenschaftlichen Fakultät der
Humboldt-Universität zu Berlin

von Dipl.-Psych. Martin Hecht

Präsident der Humboldt-Universität zu Berlin
Prof. Dr. Jan-Hendrik Olbertz

Dekan der Lebenswissenschaftlichen Fakultät
Prof. Dr. Richard Lucius

Gutachter/Gutachterin: 1. Prof. Dr. Oliver Lüdtke

2. Prof. Dr. Hans Anand Pant

3. Prof. Dr. Andreas Frey

Tag der Verteidigung: 14.07.2015

Inhaltsverzeichnis

Danksagung

Das Entstehen dieser Arbeit war ein interessanter und aufregender Prozess mit vielen Höhen und Tiefen und nur durch die Unterstützung zahlreicher Menschen möglich. Deshalb möchte ich mich ganz herzlich bei all jenen bedanken, die mich auf diesem Weg begleitet und unterstützt haben.

Diese Arbeit entstand im Rahmen meiner Tätigkeit als wissenschaftlicher Mitarbeiter am Institut zur Qualitätsentwicklung im Bildungswesen (IQB). Mein Dank gilt Prof. Dr. Hans Anand Pant und Prof. Dr. Petra Stanat für die Schaffung eines lehrreichen und stimulierenden Arbeitsumfeldes und für die Gelegenheit, am IQB promovieren zu dürfen. Das NaWi-Team ist mir im Laufe der Jahre sehr ans Herz gewachsen. Meine frequenten Aussagen, wir seien das beste Team, waren tatsächlich mehr ernst als selbstironisch gemeint!

Prof. Dr. Oliver Lüdtke möchte ich für die Betreuung meiner Dissertation und für die vielen hilfreichen Ratschläge immer genau zur richtigen Zeit danken. Prof. Dr. Hans Anand Pant hatte immer ein offenes Ohr für mich und hat durch seine herzliche und motivierende Art für einen angenehmen Verlauf meiner Promotionszeit gesorgt. Prof. Dr. Andreas Frey hat durch viele gute Hinweise und Anmerkungen entscheidend zum Gelingen von zwei der drei Publikationen beigetragen. Von Prof. Dr. Ulrich Schroeders habe ich viel über den wissenschaftlichen Prozess und Strategien zur Gestaltung wissenschaftlicher Arbeiten gelernt. Seine Begeisterung

für die Wissenschaft hat ebenfalls stark auf mich abgefärbt. Thilo Siegle stand mir unermüdlich zur Seite und hat mich immer gern beraten.

Ganz besonderen Dank möchte ich Sebastian Weirich aussprechen. In unendlichen und unendlich vielen Gesprächen haben wir unsere Dissertationen immer und immer wieder vorangetrieben. Durch ihn habe ich auch eigentlich ungenießbar starken schwarzen Tee schätzen gelernt. Ohne diese Koffeinüberdosen hätte sich die Erstellung der Dissertation wohl wesentlich länger hingezogen, oder zumindest wesentlich weniger Spaß gemacht.

Mit Stefan Schauber habe ich nicht nur regelmäßig die kulinarischen Köstlichkeiten verschiedener Essensausgabestellen auf dem Charité- und HU-Campus genossen, sondern auch an sehr spannenden Fragestellungen der medizinischen Ausbildungsforschung gearbeitet. Dr. Patricia Heitmann hat mir mit ihrem reflektierten Blick auf den Wissenschaftsprozess sehr geholfen, meine Arbeit angemessen zu bewerten und den Prozess effizient zu gestalten. Auch haben wir zusammen entschlossen und unbeirrt wichtige Bildungsforschung betrieben. Mit Steffen Zitzmann habe ich in vielen Gesprächen zahlreiche methodische Probleme ausgiebig besprochen und viele Ideen zu deren Lösung generiert. Anne Ziemke gilt mein Dank insbesondere dafür, dass sie mich hartnäckig zum Besuch von Sportkursen ermunterte beziehungsweise zwang. Robert Deutschländer half mir über die Jahre mit philosophischen Weisheiten den Blick auf das große Ganze nicht zu verlieren. Auch viele Ideen zu alternativen Lebens- und Berufswegen stammen aus spätabendlichen Sitzungen mit Robert und Mathias Deutschländer auf dem Campus; wir bleiben dran!

Die sicher wichtigste Person, der ich am meisten verdanke, war und ist Dr. Julia Wolff. Ich stehe tief in ihrer Schuld. Meinen Eltern möchte ich für die vielen Jahre der Unterstützung danken. Meiner Schwester danke ich für das gründliche Korrekturlesen. Nicht zuletzt hat mir auch Drea mit ihrer stoischen guten Laune immer wieder gezeigt, dass man das Leben genießen sollte.

Zusammenfassung

Messinstrumente stellen in der wissenschaftlichen Forschung ein wesentliches Element zur Erkenntnisgewinnung dar. Das Besondere an Messinstrumenten im Large-scale Assessment in der Bildungsforschung ist, dass diese normalerweise für jede Studie neu konstruiert werden und dass die Testteilnehmer verschiedene Versionen des Tests bekommen. Hierbei ergeben sich potentielle Gefahren für die Akkuratheit und Validität der Messung. Um solche Gefahren zu minimieren, sollten (a) die Ursachen für Verzerrungen der Messung und (b) mögliche Strategien zur Optimierung der Messinstrumente eruiert werden. Deshalb wird in der vorliegenden Dissertation spezifischen Fragestellungen im Rahmen dieser beiden Forschungsanliegen nachgegangen.

Die im Fokus stehenden Effekte beruhen auf den vielfältigen Möglichkeiten, Messinstrumente zu konstruieren. Durch unterschiedliche Zusammenstellung von Testheften können *Testhefteffekte* auftreten. Diese wurden in Daten einer Large-scale Assessment Studie nachgewiesen. Auch die Positionierung von Items innerhalb der Testhefte kann die Messung beeinflussen. Solche *Positionseffekte* ließen sich ebenfalls in empirischen Daten einer großen Schulleistungsstudie zeigen, wobei deren Größenordnung als eher klein identifiziert wurde. Neben Testheften und Itempositionen können sich auch Eigenschaften des Testdesigns auf die Messung auswirken. Durch Balancierung bestimmter Faktoren kann deren Einfluss bereits mit Hilfe des Testdesign kontrolliert werden, sodass diese Faktoren nicht im statistischen Modell berücksichtigt werden müssen. Die in den meisten Large-scale Assessment Programmen optimierten Designeigenschaften sind die *Positionsbalance* und die *Blockpaarbalance*. Mittels einer Simulationsstudie konnte gezeigt werden, dass eine höhere Positionsbalance zu einer akkurateren Parameterschätzung im Rasch-Modell führt. Hingegen zeigte sich für die Blockpaarbalance ein Nulleffekt. Für die Erstellung von Testheften sind weiterhin akkurate Aufgabenbearbeitungszeiten notwendig, damit die Sollbearbeitungszeit genau eingehalten werden kann. Deshalb wurde ein empirisch fundiertes Vorhersagemodell erstellt, mit dem die für das Testdesign unverzichtbaren Aufgabenbearbeitungszeiten aus leicht verfügbaren Aufgabeneigenschaften kostengünstig berechnet werden können.

Abstract

Measurement instruments are essential elements in the acquisition of knowledge in scientific research. Special features of measurement instruments in large-scale assessments of student achievement are their frequent reconstruction and the usage of different test versions. Here, threats for the accuracy and validity of the measurement may emerge. To minimize such threats, (a) sources for potential bias of measurement and (b) strategies to optimize measuring instruments should be explored. Therefore, the present dissertation investigates several specific topics within these two research areas.

The investigated effects arise from the manifold options in the process of constructing measurement instruments. *Booklet effects* may occur because of different compilations of booklets. Such effects were detected in a large-scale assessment study of student achievement. The position of items in booklets may influence the measurement as well. Using data from another large-scale assessment study, *position effects* were observed and identified as rather small. Besides booklets and positions, features of the booklet design might exert influences on the measurement. Balancing certain factors can help to control for biases already in the booklet design. Thus, balanced factores do not need to be included in the statistical model. Most large-scale assessment programs optimize booklet designs with respect to the *position balance* and the *cluster pair balance*. Results of a simulation study showed that a higher position balance leads to more accurate parameter estimates in the Rasch model. For cluster pair balance, a null effect occured. Another crucial issue for the construction of booklets is the availability of accurate testlet response times to ensure the accuracy of targeted booklet times. Therefore, an easy-to-use and low-cost prediction model that uses readily available testlet properties was derived from empirical data.

Liste der Beiträge

Hecht, M., Weirich, S., Siegle, T. & Frey, A. (2014). Modeling booklet effects for nonequivalent group designs in large-scale assessment. *Educational and Psychological Measurement*. Advance online publication. doi:10.1177/0013164414554219

Hecht, M., Weirich, S., Siegle, T. & Frey, A. (2015). Effects of design properties on parameter estimation in large-scale assessments. *Educational and Psychological Measurement*. Advance online publication. doi:10.1177/0013164415573311

Hecht, M., Siegle, T. & Weirich, S. (eingereicht). A model for the estimation of testlet response time to optimize test assembly in paper-and-pencil large-scale assessments. Manuskript eingereicht zur Publikation in *Large-scale Assessments in Education*.

(Stand: 09.03.2015)

Überblick über die Dissertation

Die vorliegende Dissertation ist folgendermaßen gegliedert: In der Einleitung (Kapitel 1) wird in das Themengebiet eingeführt, relevante Konzepte und Begriffe erörtert und die Forschungsvorhaben theoretisch eingebettet. Die Einleitung endet mit der Beschreibung der Ziele und Forschungsfragen dieser Dissertation (Abschnitt 1.4). In Kapitel 2 werden die Forschungsvorhaben beschrieben und die verwendeten Methoden skizziert. In Kapitel 3 erfolgt eine Zusammenfassung der Ergebnisse. Abschließend werden die Befunde diskutiert, praktische Implikationen abgeleitet, die verwendeten Methoden kritisch bewertet, die Grenzen der Arbeit ausgelotet und Ideen für zukünftige Forschungsvorhaben generiert (Kapitel 4).

In der Druckversion befinden sich die veröffentlichten und zur Veröffentlichung vorgesehenen Beiträge am Ende dieses Dokuments (Abschnitt *Beiträge* ab S. 125). Die Online-Version beinhaltet die zur Veröffentlichung vorgesehenen Beiträge, während für die veröffentlichten Beiträge die Quelle angegeben und verlinkt ist.

1 Einleitung

„Miss alles, was sich messen lässt, und mach alles messbar, was sich nicht messen lässt.“ (Galileo Galilei)

Messungen sind zentrale Bestandteile der quantitativen Wissenschaften. Durch Messungen kann die Welt kohärenter und objektiver beschrieben werden. Eine Messung kann als „the assignment of numerals to objects or events according to rules“ Stevens (1946, S. 677) definiert werden. Zum Beispiel weist eine Waage einem Objekt eine Masse zu. Diese Messung kann allerdings auf verschiedenen Prinzipien beruhen. Während eine Federwaage die auf sie einwirkende Kraft misst und aus dieser unter Berücksichtigung der Erdbeschleunigung die Masse ableitet, wird bei der Messung mit einer Balkenwaage die Gewichtskraft des zu messenden Objekts mit der Gewichtskraft einer bekannten Masse verglichen. Der Messwert ist unabhängig vom Funktionsprinzip des Messinstruments identisch: Alle Waagen messen die Masse von Objekten. Die Messgenauigkeit hingegen hängt stark von der Art der Waage und dem Kontext, in dem die Messung stattfindet, ab. Federwaagen sind auf eine bestimmte Fallbeschleunigung normiert. Deshalb ändert sich der Messwert je nach Fallbeschleunigung, die am Ort der Messung vorherrscht. Beispielsweise würde man am Nordpol einen anderen Messwert als am Äquator erhalten, da die Fallbeschleunigung an diesen beiden Orten unterschiedlich ist. Hingegen funktioniert die

Balkenwaage unabhängig vom Ort der Messung.

Während für Messungen physikalischer Größen die Kontextabhängigkeiten oft bekannt sind, und deshalb gegebenenfalls berücksichtigt (d. h. korrigiert) werden können, sind diese bei Messungen psychologischer Merkmale häufig unbekannt. Hieraus ergeben sich mehrere generelle Forschungsfragen, zu denen die vorliegende Dissertation einen Beitrag leistet:

1. Welche Kontextabhängigkeiten treten bei psychometrischen Messungen auf?

2. Wie können psychometrische Messinstrumente so konstruiert werden, dass diese kontextunabhängig funktionieren oder robust gegenüber kontextuellen Einflüssen sind?

3. Welcher Spielraum besteht bei der Konstruktion von psychometrischen Messinstrumenten?

1.1 Messungen in der Psychologie

In der Psychologie ist das zentrale Objekt, über das Erkenntnisse gewonnen werden soll, der Mensch. Eine psychologische Messung ist demnach das Zuweisen von Zahlen zu Merkmalen beziehungsweise Eigenschaften von Menschen. Die Systematik, nach denen diese Zuweisung erfolgt, ist das Messinstrument. Während physikalische Merkmale (z. B. Gewicht und Größe) durch etablierte Messinstrumente (z. B. Waage und Zollstock) bestimmt werden können, existieren für viele psychologische Merkmale keine oder nur unzureichende Messinstrumente. Dies liegt teilweise daran, dass viele psychologische Merkmale – im Gegensatz zu physischen oder physikalischen Merkmalen – nicht direkt beobachtbar sind. Während zum Beispiel das Geschlecht durch

bloßes visuelles Observieren gemessen werden kann, funktioniert diese einfache, direkte Messung bei psychologischen Merkmalen wie zum Beispiel Intelligenz oder Extraversion nicht. Vielmehr müssen solche Merkmale, die psychologisch-theoretisch konstruiert wurden und deshalb auch *Konstrukte* genannt werden, auf der Grundlage direkter Beobachtungen erschlossen werden. Zum Beispiel ist das Konstrukt Extraversion über mehrere direkt beobachtbare Merkmale definiert: „The typical extravert is sociable, likes parties, has many friends, needs to have people to talk to, and does not like reading or studying by himself." (Eysenck & Eysenck, 1975, S. 5). Diese direkt beobachtbaren Verhaltensweisen, Einstellungen und Eigenschaften dienen als Indikatoren für das nicht direkt beobachtbare *latente* Konstrukt. Konkret könnte man die Affinität, auf Partys zu gehen, die Anzahl an Freunden und die Frequenz des Lesens von Büchern erfragen (oder durch Beobachtungen ermitteln), und mit Hilfe dieser Informationen einer Person einen numerischen Wert auf dem Konstrukt *Extraversion* zuweisen. Die Art und Weise, wie die Indikatoren zum latenten Konstrukt mathematisch in Verbindung stehen, wird hierbei durch das Messmodell festgelegt.

In der Psychologie existieren primär zwei generelle Ansätze („Schulen"), die sich mit der Entwicklung von Messmodellen befassen: die klassische Testtheorie (KTT) und die Item-Response-Theorie (IRT)[1]. In beiden Ansätzen wird versucht, numerische Werte auf den Indikatoren geeignet zu Werten auf dem latenten Konstrukt zuzuordnen. Der Hauptunterschied besteht darin, dass für die manifesten Indikatoren in der KTT Intervallskalenniveau, und in der IRT Nominal- oder Ordinalskalenniveau (Stevens, 1946) angenommen wird. Unterschiedliche Skalenniveaus zeichnen sich durch einen unterschiedlichen Informationsgehalt aus. Werte auf Nominalskalen

[1]Im deutschsprachigen Raum wird statt IRT auch der Begriff *Probabilistische Testtheorie* verwendet. Dieser Begriff wird jedoch von Steyer und Eid (2001) als missverständlich bewertet, da sowohl KTT- als auch IRT-Modelle stochastische Messmodelle sind.

können nur dahingehend interpretiert werden, dass jene sich unterscheiden. Zum Beispiel ist die Zugehörigkeit zu bestimmten Gruppen (z. B. Männer/Frauen, Parteien) nominal. Es gibt keine Distanz zwischen den Kategorien und auch keine Ordnung. Werte auf Ordinalskalen besitzen hingegen genau diese Ordnung, was Aussagen darüber ermöglicht, welcher Wert größer als ein anderer ist. Ein Beispiel ist die Rangfolge in einem Laufwettbewerb. Hierbei weiß man, wer früher ins Ziel gekommen ist, aber nicht wie viel früher. Diese zusätzliche Information bieten erst Intervallskalen. Obwohl sich die KTT und die IRT im Skalenniveau der beobachteten Variablen unterscheiden, wird dennoch das gleiche Ziel der Messung von latenten Konstrukten verfolgt. Trotzdem haben sich diese beiden Ansätze relativ unabhängig voneinander entwickelt. Auf dem Gebiet der Schulleistungsforschung waren und sind Modelle der IRT dominant. Deshalb wird in dieser Dissertation auf IRT-Modelle fokussiert.

1.2 Messmodelle in der Schulleistungsforschung

1.2.1 Das Rasch-Modell

In der Schulleistungsforschung stellt die Leistung[2] der Schülerinnen und Schüler in einem bestimmten Fach oder Themengebiet das zu messende latente Konstrukt dar. Als manifeste Indikatoren für die latente Leistung dienen hierbei meist dichotome Items, die – im Gegensatz zu Einstellungs- oder Persönlichkeitsitems – entweder richtig (meist als 1 kodiert) oder falsch (meist als 0 kodiert) beantwortet werden können. Aus diesem Grund wird in der Schulleistungsforschung besonders auf IRT-Methoden zurückgegriffen. Eines der populärsten Messmodelle im IRT-Kontext ist das Rasch-Modell (Rasch, 1960). In diesem wird die Wahrscheinlichkeit P, dass ein

[2] In dieser Arbeit werden die Begriffe *Leistung*, *Fähigkeit* und *Kompetenz* gleichermaßen und ohne theoretische Differenzierung verwendet, um die zu messende Personenvariable zu benennen.

Schüler j ein Item i löst, auf einen Personenparameter θ_j (interpretierbar als die Fähigkeit des Schülers) und einen Itemparameter β_i (interpretierbar als die Schwierigkeit des Items) bedingt:

$$P(X_{ji} = 1) = \text{logit}^{-1}(\theta_j - \beta_i) \tag{1.1}$$

Die Werte θ_j und β_i auf der latenten Skala werden als reelle Zahlen im Bereich [-inf,+inf] angenommen. Da sich Wahrscheinlichkeiten allerdings nur im Wertebereich von [0,1] bewegen, muss eine geeignete Transformation vorgenommen werden, damit dieser Wertebereich eingehalten wird. Diese Transformation wird im Rasch-Modell durch das Logarithmieren des Quotienten aus Lösungswahrscheinlichkeit und Falschbeantwortungswahrscheinlichkeit („Wettquotient“) vorgenommen:

$$\text{logit}(P(X_{ji} = 1)) = \log\left(\frac{P(X_{ji} = 1)}{1 - P(X_{ji} = 1)}\right) \tag{1.2}$$

Die Werte auf der latenten Skala werden aus diesem Grund auch „Logits“ genannt. Als mathematisch äquivalente Formulierung können statt der Logit-Transformation der Lösungswahrscheinlichkeit auch die Werte der latenten Skala durch die inverse Logit-Funktion logit^{-1} wie in Gleichung 1.1 transformiert werden (die hochgestellte „-1“ kennzeichnet, dass es sich um die *inverse* Logit-Funktion $\text{logit}^{-1}(\alpha) = \frac{1}{1+e^{-\alpha}} = \frac{e^{\alpha}}{e^{\alpha}+1}$ handelt). Mit den bisherigen Darlegungen ist die Verlinkung der Lösungswahrscheinlichkeit mit den additiv verknüpften Effekten (Personenfähigkeit und Itemschwierigkeit) beschrieben. Zu einer vollständigen Formulierung des Rasch-Modells gehört weiterhin aber noch eine Verteilungsannahme der Responses auf den Items. Da diese nur zwei Werte annehmen (ein Item ist entweder gelöst oder nicht)

wird eine Bernoulli-Verteilung mit dem Lageparameter $P(X_{ji} = 1)$ angenommen:

$$X_{ji} \sim \text{Bernoulli}(P(X_{ji} = 1)) \tag{1.3}$$

Das Rasch-Modell ist durch die drei Komponenten Modellgleichung (Gleichung 1.1), der Logit-Funktion als Link-Funktion (Gleichung 1.2) und der Verteilungsannahme (Gleichung 1.3) konzeptuell vollständig spezifiziert. Weiterhin können aber auch zusätzliche Annahmen eingeführt werden, zum Beispiel Verteilungsannahmen über die Personen- und Itemparameter. Aus der Modellgleichung wird erkenntlich, dass für jede Person $1, \ldots, J$ und jedes Item $1, \ldots, I$ ein Parameter existiert. Diese Parameter werden auch *fixed effects* genannt. Damit müssen $J + I$ Modellparameter geschätzt werden. Da mit steigender Anzahl an Personen und Items die Anzahl an Modellparametern mit der gleichen Rate steigt, sind die Parameterschätzungen nicht konsistent (Neyman & Scott, 1948). Das heißt, dass diese mit steigender Anzahl an Datenpunkten nicht gegen ihren wahren Wert konvergieren. Da inkonsistente Parameterschätzungen die Güte und Aussagekraft eines Messmodells gefährden (Tuerlinckx et al., 2004), kann zur Vermeidung dieses Problems eine Verteilungsannahme über die Personen- und/oder Itemparameter getroffen werden. Damit reduziert sich die Anzahl der Modellparameter auf einen Lageparameter (Mittelwert) und einen Dispersionsparameter (Varianz) der jeweiligen Verteilung. Damit ergeben sich vier Versionen des Rasch-Modells in Abhängigkeit der Modellierung von Personen und Items (De Boeck, 2008): (a) für jede Person und jedes Item jeweils ein Parameter (fixed persons–fixed items, FPFI), (b) Personen und Items jeweils mit Verteilungsannahme (random persons–random items, RPRI), (c) Personen mit Verteilungsannahme und Items als fixed effects (RPFI) und (d) Personen als fixed effects und Items mit Verteilungsannahme (FPRI). Zusätzlich zur technischen Komponente

besitzt die Einführung von Verteilungsannahmen auch konzeptuelle Implikationen. Während bei der Modellierung als fixed effects die Schätzung von Parametern für konkrete Personen und Items im Fokus steht, werden die Personen und Items bei der Modellierung als Zufallsvariable als eine Ziehung aus einer Population und damit prinzipiell als austauschbar angesehen.

Die Wahl der Rasch-Modell-Variante wird durch das spezifische Anwendungsproblem und die konkrete Forschungsfrage bestimmt. Würde man zum Beispiel einen Test entwickeln und an einer Population normieren wollen, sollte das RPFI-Rasch-Modell verwendet werden, da die Itemschwierigkeiten in der Normpopulation die entscheidenden Referenzgrößen sind, die später verwendet werden können, um die Kompetenzen von Schülerinnen und Schülern einer anderen Population in Relation zur Normpopulation darzustellen. Für diagnostische Zwecke sind hingegen die Fähigkeiten einzelner Personen relevant. Deshalb sollte hierzu das FPRI-Rasch-Modell zur Anwendung kommen, da in diesem Modell für jede Person ein Parameter geschätzt wird. Für bestimmte Forschungsfragen (z. B. der Modellierung von Kontexteffekten), bei denen für Personenfähigkeiten und Itemschwierigkeiten nur „kontrolliert“ werden soll, bietet es sich an, dass RPRI-Rasch-Modell als Basis zu verwenden und entsprechend zu erweitern. Dieses Vorgehen wird auch in dieser Dissertation zur Modellierung spezieller Kontexteffekte (Testhefteffekte im Abschnitt 2.1.1; Positionseffekte im Abschnitt 2.1.2) angewandt.

Die Popularität des Rasch-Modells begründet sich auf dessen Plausibilität, Einfachheit sowie speziellen Eigenschaften, die bestimmte Interpretationen zulassen (z. B. Rost, 2004). Durch die additive Verknüpfung von Item- und Personenparametern wird eine Trennung von deren Einflüsse ermöglicht. Es kommt demnach nicht darauf an, welche Items von welcher Person gelöst wurden, sondern nur darauf, wie oft ein Item gelöst wurde beziehungsweise wie viele Items eine Person gelöst

hat (*suffiziente Statistiken*). Eine weitere günstige Eigenschaft des Rasch-Modell ist die Unabhängigkeit der Differenz zweier Personenparameter von der Verteilung der Itemparameter beziehungsweise die Unabhängigkeit der Differenz zweier Itemparameter von der Verteilung der Personenparameter (*spezifische Objektivität*). Dadurch werden Vergleiche von Personen (bzw. Items) möglich, die von den Items (bzw. Personen) unabhängig sind.

Aufgrund der vorteilhaften Eigenschaften des Rasch-Modells wird in einigen großen Programmen der Bildungsforschung (z. B. PISA, OECD, 2012; Bildungsstandards in Deutschland, Pant et al., 2013) versucht, Tests zu entwickeln, für die das Rasch-Modell gilt. Hierfür werden Items in mehreren Phasen entwickelt, empirisch getestet und so überarbeitet, dass diese optimal in das Rasch-Modell passen. Dieser aufwändige Entwicklungsprozess ist notwendig, da Tests, die nicht speziell in Richtung Rasch-Modell entwickelt wurden, normalerweise auch nicht inhärent dem Rasch-Modell genügen. Als Alternativen zum Rasch-Modell existieren eine große Anzahl verschiedener Modelle, die meist aber auch sowohl Personenfähigkeits- als auch Itemschwierigkeitsparameter beinhalten, und somit als Erweiterungen des Rasch-Modells verstanden werden können.

1.2.2 Erweiterungen des Rasch-Modells

Die im Raschmodell getroffene Annahme gleicher Itemtrennschärfen kann durch Einführung eines itemspezifischen Trennschärfeparameters, der multiplikativ mit dem Personenparameter und dem Itemschwierigkeitsparameter verknüpft ist, relaxiert werden. Dieses von Birnbaum (1968) vorgeschlagene 2PL-Modell hat einerseits zur Folge, dass Schätzverfahren für nichtlineare Modelle angewandt werden müssen, da die Modellparameter nichtlinear (d. h. nicht additiv) verbunden sind. Anderseits

kommt es jetzt nicht nur darauf an, wie viele Items eine Person gelöst hat, sondern auch noch welche Items. Der itemspezifische Trennschärfeparameter stellt somit ein Gewichtungsfaktor für die Relevanz eines Items bei der Messung des latenten Konstrukts dar. Neben den Itemtrennschärfen können auch noch weitere Parameter eingeführt werden, wie beispielsweise ein Rateparameter (3PL-Modell) oder ein „Slippingparameter“ (4PL-Modell, siehe z. B. De Ayala, 2009).

Neben Personen- und Itemparametern sind weiterhin eine Vielzahl an anderen Effekten denkbar, die potentiell einen Einfluss auf die Lösungswahrscheinlichkeit haben können. Zwei klassische Beispiele sind *Positionseffekte* und *Reihenfolgeeffekte* (z. B. Rost, 2004). Da zur Messung von latenten Konstrukten immer mehrere Items verwendet werden, ist eine bestimmte Reihenfolge der Items im Test unvermeidlich. Dadurch werden verschiedene Items an verschiedenen Positionen im Test präsentiert. Wenn diese Positionierung von Items die Lösungswahrscheinlichkeit beeinflusst, spricht man von Positionseffekten. Das Zustandekommen solcher Effekte wird meist auf Ermüdung, schwindende Testmotivation und Zeitmangel am Testende oder auf mangelndes Instruktionsverständnis am Testanfang zurückgeführt (Rost, 2004). Reihenfolgeeffekte entstehen hingegen dadurch, dass die Beantwortung eines Items davon abhängt, welche Items davor beantwortet wurden. Zum Beispiel sind bei vielen Leistungs- und Intelligenztests die Aufgaben nach steigender Schwierigkeit geordnet, was zur Folge haben kann, dass die schwierigen Aufgaben infolge der Übung an leichteren Aufgaben ebenfalls etwas leichter zu lösen sind (Rost, 2004).

Zeitmangel als Ursache für Positionseffekte stellt ein besonderes Problem in Leistungstests dar. Wenn die zu messende Leistung nicht darin besteht, bestimmte Aufgaben so schnell wie möglich zu lösen (*speed test*), sondern das zu messende Konstrukt als bearbeitungszeitunabhängig konzipiert ist (*power test*), beeinträchtigt eine zu rigide Zeitvorgabe die Validität des Tests (Lu & Sireci, 2007). Während zum Bei-

spiel Zeitvorgaben bei Tests, die die Tippgeschwindigkeit von Schreibkräften messen, zur Validität beitragen, sinkt bei Leseverstehenstests die Validität bei zu hohem Zeitdruck, da die Lesegeschwindigkeit nicht konstruktinhärent ist, aber trotzdem teilweise mitgemessen wird. Für die Validität von Power Tests wäre es demnach am sinnvollsten, keine Bearbeitungzeitvorgaben zu machen und allen Testteilnehmern so viel Zeit wie benötigt zu geben. Dieses Vorgehen ist in Schulleistungsstudien allerdings nicht praktikabel, da aus administrativen Gründen meist Schülerinnen und Schüler vor Ort (in Schulen) und dadurch in genormten Zeiteinheiten (Schulstunden) getestet werden.

Die zur Testung zur Verfügung stehende Zeit ist also begrenzt und wird vorab extern (z. B. durch die Schuladministration) definiert. Als beeinflussbare Größe für die Zusammenstellung des Tests bleibt aber die Anzahl an Items. Hier stellt sich die Frage, wie hoch die optimale Anzahl an Items zur Messung eines Konstrukts unter Berücksichtigung der vorgegebenen Testzeit ist. Einerseits erhöht sich die Reliabilität der Messung mit steigender Itemanzahl. Andererseits verstärken sich Positionseffekte bei hohem Zeitdruck und die Validität des Power Tests verringert sich durch unintendiertes Mitmessen einer Speed-Komponente. Bezüglich dieser beiden antinomischen Effekte muss ein Kompromiss im Sinne eines plausiblen Optimums gefunden werden. Dieser Abwägungsprozess läuft letzten Endes auf die Festlegung hinaus, wie hoch die Rate an Schülerinnen und Schülern ist, die den Test komplett schaffen. In PISA (OECD, 2005) und den IQB-Studien zu den Bildungsstandards in den Naturwissenschaften wurde diese Rate auf 90% festgelegt. Das heißt, im Test werden so viele Items untergebracht, dass 90% der Schülerinnen und Schüler in der vorgegebenen Testzeit den Test komplett bearbeiten können. Unabhängig davon, auf welchen Wert man diese Sollkomplettierungsrate festlegt, ist es von großer Wichtigkeit, die exakte Bearbeitungszeit von Items zu kennen. Nur dann können Tests zusammengestellt werden, die eine bestimmte Sollbearbeitungszeit besitzen.

1.3 Testdesigns und Kontexteffekte im Large-scale Assessment

„Although proper examination of the results of an experiment is important, there is no way that a clever analysis can make up for a poorly designed study, a study that leaves out key factors, or inadvertently confounds and/or masks relevant factors.“ (Giesbrecht & Gumpertz, 2004, S. 2)

Die richtige Wahl von angemessenen statistischen Auswertungsmethoden und Messmodellen ist nur eine Komponente, die zum Erfolg von Large-scale Assessments beiträgt. Ebenso wichtig ist ein passendes Testdesign. Suboptimale Eigenschaften von Testdesigns und Fehler bei der Erstellung des Testdesigns können nach der Datenerhebung nicht – oder nur mit hohem Aufwand – innerhalb der Datenauswertung korrigiert werden. Es ist daher von hoher Wichtigkeit, Testdesigns mit Hinblick auf die Ziele und Forschungsfragen der jeweiligen Studie zu optimieren.

Da instrumentenseitige Kontexteffekte (*instrument effects*, Brennan, 1992) im engen Zusammenhang mit der Erstellung von Testdesigns stehen, werden diese beiden sich bedingenden Themen in diesem Abschnitt gemeinsam behandelt. Als erstes wird definiert, über welche Einheiten Aussagen in Large-scale Assessment Studien getroffen werden sollen, und eruiert, ob der Einsatz mehrerer Testformen vertretbar ist. Danach wird die Idee des *Multiple Matix Sampling* beschrieben, das den meisten Testdesigns im Large-scale Assessment zugrunde liegt. Anschließend werden Kontexteffekte theoretisch eingeführt und Strategien zur Handhabung von Kontexteffekten erörtert, wobei ein besonderer Fokus auf den in der vorliegenden Dissertation thematisierten instrument effects liegt. Im Anschluss wird die Erstellung von Testdesigns und den dabei resultierenden Designeigenschaften beschrieben. Abschließend

werden verschiedene Ursachen für das Auftreten von Verzerrungen (Bias), speziell auch im Rahmen von IRT-Modellen, diskutiert.

1.3.1 Eine oder mehrere Testformen?

Tests zur Messung der Leistung einzelner Personen zu individualdiagnostischen Zwecken bestehen aus einer festgelegten Menge an Items, die allen Personen in der gleichen Reihenfolge zur Bearbeitung vorgelegt werden. Es existiert also genau eine Version des Tests, eine *Testform*. Dies ist sehr vorteilhaft, um relevanten Personenkreisen, die nicht unmittelbar in die Testung involviert sind, aber trotzdem von Folgen individualdiagnostischer Entscheidungen betroffen sind, keinen Anlass zu geben, an der Fairness des Tests zu zweifeln. Zum Beispiel bekommen Schülerinnen und Schüler, die am Ende der Grundschule keine Empfehlung für das Gymnasium erhalten haben, die Möglichkeit, einen Eignungstest zu bearbeiten, um die Erlaubnis für den Gymnasialbesuch doch noch zu erlangen. Würde man diesen Schülerinnen und Schülern verschiedene Aufgaben geben oder dieselben Aufgaben in verschiedener Reihenfolge, gäbe man Eltern einen Grund, die Testergebnisse infrage zu stellen. Da der Gymnasialbesuch für Eltern, besonders für diejenigen, die ihre Kinder zum Eignungstest schicken, eine hohe Wertigkeit besitzt, wäre hierbei durchaus mit größeren Problemen bis hin zu rechtlichen Auseinandersetzungen zu rechnen. Für Schulleistungstests, die nicht auf die Messung individueller Kennwerte, sondern aggregierter Gruppenkennwerte, abzielen, ist die Notwendigkeit einer einzigen Testform nicht mehr gegeben. Solche Studien, in denen die Ermittlung von Kompetenzcharakteristiken auf Gruppenebene im Fokus steht, werden aufgrund ihrer meist vergleichsweise großen Stichproben *Large-scale Assessments* genannt. Da sich die Samplingeinheiten (Schülerinnen und Schüler) und die Analyseeinheiten (Gruppen von Schülerinnen und Schülern) also unterscheiden, müssen die Messungen in solchen Studien

auch nicht mehr auf Individualebene fair und korrekt sein. Es können demnach auch mehrere Testformen eingesetzt werden, da nicht mehr für diejenigen Einheiten (Individuen) Vergleiche vorgenommen werden, die unterschiedliche Testformen bekommen.

1.3.2 Multiple Matrix Sampling

Die Verwendung mehrerer, nicht identischer Testformen bietet weitere Vorteile: Jede Testform muss nicht alle Items enthalten, sondern lediglich aus einer kleinen Teilmenge des gesamten *Itempools* bestehen. Dadurch kann zum einen eine große Itemmenge (auch mit Items verschiedener Domänen) in einer Large-scale Assessment Studie untergebracht werden. Zum anderen kann trotz einer großen Gesamtmenge an Items die Bearbeitungszeit je Testteilnehmer in zumutbaren Grenzen gehalten werden. Das Problem der begrenzten Verfügbarkeit von Testzeit hatte bereits Lord (1962, S. 259) erkannt: „[t]he most serious obstacle is the fact that not every school is willing ... to ... require its students to spend a class period or more taking tests“, und führt als Problemlösung an: „The problem of getting each school's cooperation would be less serious if only a few moments of each student's time were required, rather than an entire class period“.

Die Methode der Verteilung von Items auf verschiedene Testformen ging in die Literatur unter dem Begriff *item sampling* (Johnson & Lord, 1958; Lord, 1962), später auch *multiple matrix sampling* (Shoemaker, 1971a), ein. Die offensichtlichen Vorteile des Multiple Matrix Samplings verhalfen dieser Methode zu großer Popularität und lösten große Forschungsaktivität aus (z. B. Plumlee, 1964; Owens & Stufflebeam, 1969; Shoemaker, 1970a; Shoemaker, 1970b; Pugh, 1971; Shoemaker, 1971b; Shoemaker, 1971c; Barcikowski, 1972; Moy, 1973; Barcikowski, 1974; Feldt

& Forsyth, 1974; Myerberg, 1975; Scheetz & Forsyth, 1977; Myerberg, 1979; Gressard & Loyd, 1991; Childs & Jaciw, 2003). Ingesamt legen die Ergebnisse dieser Forschung nahe, dass Multiple Matrix Sampling mindestens genauso adäquate oder bessere Schätzungen für Gruppenkennwerte als klassische Designs mit einer Testform liefern. Johnson & Lord (1958, S. 328) reizen die Idee des Multiple Matrix Sampling Ansatzes sogar vollständig aus: „In theory, a national mean for a complete test could be obtained by administering only one item to each student“. Dieser Vorschlag mag zwar zu statistisch korrekten Ergebnissen führen, ist in der Praxis aber eher ungünstig, da es aus testadministrativen Gründen unökonomisch ist, Schülerinnen und Schülern nur ein Item vorzulegen. Wenn man erst einmal alle Genehmigungen eingeholt hat und Testleiterinnen und -leiter in die Schulen schickt, ist es sinnvoller, gleich Tests im Umfang der Zumutbarkeit (meist ca. eine oder wenige Schulstunden) bearbeiten zu lassen.

1.3.3 Kontexteffekte und deren theoretische Einbettung

Multiple Matrix Designs wurden aufgrund besagter Vorteile später auch in Large-scale Assessment Programmen in der Bildungsforschung eingesetzt, zum Beispiel im *National Assessment of Educational Progress* (NAEP; Beaton, 1987; Beaton, 1988a). Allerdings zeigten sich schnell auch Nachteile dieser Designs. Von einer Erhebung (1984) zur nächsten (1986) sank die Leseleistung der getesteten 17-Jährigen stark („a surprisingly large decrease“, Zwick, 1991, S. 11), nachdem diese über viele Jahre sehr stabil war. Dieser Effekt ging in die Literatur als *1986 NAEP reading anomaly* ein (Beaton, 1988b; Haertel, 1989; Beaton & Zwick, 1990). Als Ursache für diesen zunächst rätselhaften Effekt wurde später das Auftreten von Kontexteffekten identifiziert (Zwick, 1991). Die Leseitems waren im Testdesign von 1984 mit Schreibitems kombiniert; 1986 hingegen mit Items aus der Mathematik und den Naturwissenschaften. Weiterhin wurden die Zusammenstellung, die Reihenfolge und die Zeitvorgaben der Leseitems verändert. Dies zeigt, dass bei der Verwendung von Multiple Matrix Designs durchaus Vorsicht geboten ist. Tatsächlich hatte Lord (1962) bereits vor solchen Kontexteffekten gewarnt: „Any practical application of the item-sampling method thus involves the further assumption that the examinees' performance on the items is not too greatly affected by the context in which they are administered“ (S. 263).

Unter der Bezeichnung *Kontexteffekte* wurden bereits eine Vielzahl von Phänomenen untersucht (siehe die umfassende Literaturanalyse von Leary und Dorans, 1985). Allerdings beklagt Brennan (1992), dass die Literatur zu Kontexteffekten quasi komplett frei von Definitionen ist. Stattdessen wird jedwede Abweichung eines Kennwerts in verschiedenen Situationen als Evidenz für das Wirken von Kontexteffekten gedeutet. Brennan stellt weiterhin fest, dass kein kohäsiver theoretischer Rahmen existiert, um die Existenz, die Größe und die Konsequenzen von Kontexteffekten

systematisch zu untersuchen und schlägt deshalb unter dem Namen *heuristic modeling* mehrere Konzepte und Denkansätze vor, die bei der theoretischen Erörterung von Kontexteffekten helfen könnten.

Als Hilfe zur Einordnung von Kontexteffekten gibt Brennan folgende Formel (Gleichung 1 auf S. 238) an:

$$\begin{aligned} \text{observed score} \quad = \quad & \text{grand mean} + \\ & \text{person effect} + \\ & \text{instrument effect} + \\ & \text{conditions effect} + \\ & \text{instrument–by–conditions interaction effect} + \\ & \text{person–by–instrument interaction effect} + \\ & \text{person–by–conditions interaction effect} + \\ & \text{residual effect} \end{aligned} \tag{1.4}$$

Der beobachtete Wert ist demnach auf Effekte der Person, des Instruments, der Bedingungen, unter denen die Messung stattfindet, und deren Interaktionen zurückzuführen. Da im Large-scale Assessment die Effekte der Person die interessierenden Messwerte darstellen, sind alle anderen Effekte potentielle „Störgrößen“, die im Kontext der Messung auftreten können.

Neben der Klassifizierung von Kontexteffekten wird in Brennans Theorie zwischen zwei Universen unterschieden: (1) *universe of generalization* und (2) *universe of allowable observations*. Ersteres ist das validitätsdefinierende Universum, während das zweite eine restringierte Version des ersteren ist. Kontexteffekte entstehen (teilweise) dadurch, dass durch Standardisierung einer Messung ein Universum impliziert wird, das sich in einer gewissen Art und Weise systematisch vom uni-

verse of generalization unterscheidet. Erfolgt eine Messung also in standardisierten Bedingungen, kann das universe of allowable observations als Teil des universe of generalization angesehen werden, und zwar als jenen Teil, in dem die standardisierten Bedingungen nur ein Set aller möglichen Bedingungen sind. Wenn über ein größeres Set an Bedingungen generalisiert werden soll, können Inferenzen basierend auf den Ergebnissen der standardisierten Messung durch Kontexteffekte beeinflusst sein. Wird beispielsweise ein Test immer nur zu einer bestimmten Uhrzeit durchgeführt, können die gewonnenen Ergebnisse theoretisch nicht auf Messungen zu anderen Uhrzeiten generalisiert werden. Es müsste erst gezeigt werden, dass Messungen zu verschiedenen Uhrzeiten äquivalent sind; ansonsten läge ein *conditions effect* vor. Analog verhält es sich für die Konstruktion des Messinstruments. Werden Items in einem Test immer in der gleichen Reihenfolge präsentiert, gelten die Ergebnisse nur für diese Itemreihenfolge. Verändert eine andere Itemreihenfolge die Messergebnisse liegt ein Kontexteffekt aus der Kategorie *instrument effect* vor. Auch Interaktionen von Personen mit dem Instrument und den Bedingungen können Kontexteffekte hervorrufen: Demotiviert eine bestimmte Zusammenstellung an Items schlechte Testteilnehmer besonders, weil diese aus vielen schweren Items besteht, liegt ein *person–by–instrument interaction effect* vor. Als *person–by–conditions interaction effect* zu interpretieren wäre, wenn melancholische Testteilnehmer an Regentagen besonders demotiviert sind und deshalb noch schlechter abschneiden.

In Brennans Konzeption liegen Kontexteffekte also immer dann vor, wenn Effekte des Instruments, der Messbedingungen oder Interaktionen von Personen mit diesen im universe of allowable observations von denen im universe of generalizations abweichen, also wenn die Generalisierbarkeit der Messergebnisse eingeschränkt ist. Aus dieser Definition lässt sich auch ableiten, wann keine Kontexteffekte vorliegen, nämlich dann, wenn entweder der Effekt in beiden (nicht identischen) Universen gleich

wäre oder wenn die beiden Universen identisch wären. Entspräche beispielsweise der Effekt einer bestimmten Itemreihenfolge jenen aller möglichen Itemreihenfolgen – gäbe es also keine Unterschiede zwischen verschiedenen Itemreihenfolgen – liegt kein Kontexteffekt vor. Ebenfalls kann kein Kontexteffekt vorliegen, wenn es theoretisch gar nicht die Möglichkeit gäbe, einen Faktor zu variieren. In diesem Fall sind die beiden Universen identisch. Beispielsweise kann in einem Test, der nur aus einem Item besteht, die Position des Items nicht variiert werden. Dementsprechend kann auch kein Positionseffekt vorliegen.

Insgesamt bietet Brennans Framework einen soliden theoretischen Rahmen, mit dessen Hilfe in spezifischen Anwendungsfällen eruiert werden kann, ob ein Kontexteffekt vorliegt oder nicht. Die theoretische Einbettung der in dieser Dissertation untersuchten Kontexteffekte erfolgt in den Abschnitten 2.1.1 (Testhefteffekte), 2.1.2 (Positionseffekte) und 2.1.3 (Designeffekte).

1.3.4 Strategien zur Handhabung von Kontexteffekten

Zur Handhabung von Kontexteffekten existieren verschiedene Strategien, die auf verschiedenen konzeptuellen Ebenen ansetzen. Einerseits können Kontexteffekte bei der statistischen Auswertung modellbasiert berücksichtigt und dadurch kontrolliert werden (*Modellierung*). Andererseits kann Kontexteffekten bereits im Testdesign durch die Strategien der *Randomisierung*, *Balancierung* und *Konstanthaltung* begegnet werden. Diese Methoden sind aus der Literatur zu experimentellen Designs (z. B. Giesbrecht & Gumpertz, 2004) entlehnt und werden auch von Yousfi & Böhme (2012) im Rahmen von Testdesigns beschrieben.

Bei der Modellierung von Kontexteffekten werden Variablen, die einen potentiellen Einfluss auf die Messung ausüben können, in das Modell integriert. Damit wird sichergestellt, dass die Modellparameter nicht durch Missspezifikation des Modells verzerrt werden (siehe auch Abschnitt 1.3.6). Im Prinzip sollte ein Messmodell deshalb alle erdenklichen Einflussfaktoren enthalten. Dies hat allerdings folgende Nachteile:

1. Die Modelle werden komplexer. Dadurch steigt die Gefahr von Fehlspezifikationen und Fehlinterpretationen.

2. Die Rechenzeit steigt, da eine größere Anzahl an Parametern beruhend auf einer größeren Datenmenge geschätzt werden muss.

3. Die Auswahl an Softwarepaketen ist geringer, da nicht in allen Softwares alle Modelle geschätzt werden können.

4. Die Schätzung der Kontexteffekte selbst kann inakkurat sein, wenn das Design nicht zu deren Schätzung optimiert wurde (Weirich, Hecht & Böhme, 2014).

5. Die Kontexteffekte sind normalerweise inhaltlich uninteressant. (Außer man möchte diese zum Beispiel im Rahmen einer Dissertation untersuchen.)

Deshalb ist es ratsam, bereits im Testdesign – und somit *vor* der Datenerhebung – Maßnahmen zur Handhabung von Kontexteffekten zu ergreifen.

Eine in experimentellen Designs häufig verwendete Methode zur Kontrolle von Drittvariablen ist die Randomisierung. Hierbei werden Einheiten *zufällig* zu verschiedenen Bedingungen zugewiesen, um die Wahrscheinlichkeit zu verringern, dass Eigenschaften der Einheiten die Messung verzerren. Da im Large-scale Assessment auch das Messinstrument durch einen Sampling-Prozess erstellt wird (Multiple Matrix Sampling, siehe Abschnitt 1.3.2), könnte das Prinzip der Randomisierung verwendet werden, um für Einflüsse verschiedener Eigenschaften des Messinstruments auszugleichen. Hierzu müssten Testhefte aus der Population aller möglichen Testhefte zufällig gezogen werden. Jedoch kann erst bei einer hinreichend großen Anzahl an Testheften davon ausgegangen werden, dass sich Testhefteffekte nivellieren. Hier liegt allerdings das Problem der randomisierten Testhefterstellung, da aus Kostengründen die Anzahl an Testheften meist niedrig gehalten werden muss. Deshalb ist die Strategie, Testhefte durch randomisierte Zuordnung von Items zu erstellen, eher ungeeignet. In anderen Anwendungsfällen ist die Randomisierung jedoch durchaus geeignet, zum Beispiel bei der Zuweisung von Testheften zu Testteilnehmern (siehe Abschnitt 2.2.1).

Bei der Methode der Balancierung wird ebenfalls versucht sicherzustellen, dass Beobachtungen im Mittel gleich stark vom jeweiligen Kontexteffekt beeinflusst werden. Im Gegensatz zur Randomisierung muss hierbei der Kontexteffekt aber bekannt sein, damit die Ausprägungen des entsprechenden Kontextfaktors systematisch so zu den Beobachtungen verteilt werden können, dass sich deren Einfluss ausmitteln kann. Balancieren bietet sich vor allem auch bei Kontextfaktoren an, die nicht kon-

stant gehalten werden können. Das prominenteste Beispiel im Large-scale Assessment Kontext für derartige Faktoren ist die Positionierung von Items im Testheft. Da aus ökonomischen Gründen jeder Testteilnehmer normalerweise mehr als ein Item bearbeitet, ist eine Reihenfolge der Items unausweichlich. Wenn die Lösungswahrscheinlichkeit eines Items davon abhängt, an welcher Position das Item präsentiert wird, spricht man von Positionseffekten. Um diese Positionseffekte auszumitteln, kann man das Auftreten eines Items balancieren, indem das Item (bzw. der Block, zu dem das Item gehört) an allen Positionen gleich häufig platziert wird. Dadurch ist zwar jedes Item immer noch von Positionseffekten betroffen, alle Items im Mittel jedoch gleich stark.

Um Kontexteffekte zu handhaben, kann die Messung auch durch Konstanthaltung von Kontextfaktoren standardisiert werden. Dadurch wird deren Einfluss zwar nicht eliminiert; da der Kontextfaktor aber für alle Beobachtungen konstant ist, ist dessen Einfluss auf die Messung ebenfalls konstant. Somit sind die Messergebnisse vergleichbar. Durch Konstanthaltung wird allerdings deren Generalisierbarkeit eingeschränkt. Es können dann keine Aussagen mehr darüber gemacht werden, wie die Ergebnisse unter anderen Gegebenheiten sind.

Zusammenfassend kann festgehalten werden, dass zur Handhabung von Kontexteffekten die vier Strategien der Modellierung, Randomisierung, Balancierung und Konstanthaltung existieren. Diese weisen unterschiedliche Vor- und Nachteile auf. Statt eine Modellierung von Kontexteffekten nach der Datenerhebung anzustreben, bietet es sich bereits an, Kontexteffekte proaktiv bei der Planung der Messung zu berücksichtigen. Dazu gehört vor allem die Erstellung des Testdesigns, die im nächsten Abschnitt beschrieben wird.

1.3.5 Erstellung und Eigenschaften von Testdesigns

In Testdesigns im Large-scale Assessment sind die Items die basale Einheit. Zur besseren Handhabung werden diese jedoch zu *Blöcken* (engl. *cluster*) gruppiert, wobei jedes Item genau einem Block zugeordnet wird, womit die Blöcke disjunkt sind. Die Bildung von Blöcken kann nach verschiedenen Kriterien erfolgen. Zum Beispiel können Items des gleichen Kompetenzbereiches zusammengefasst werden. Ein wichtiger Vorteil der Bildung von Blöcken ist vor allem aber die Konstruktion von Einheiten mit gleicher Bearbeitungszeit. Während sich die Bearbeitungszeiten von Items normalerweise unterscheiden, werden die Blöcke so zusammengestellt, dass alle dieselbe Bearbeitungszeit (z. B. 20 Minuten) besitzen. Erst so werden die Blöcke im Testdesign austauschbar, da somit gewährleistet ist, dass die Gesamtbearbeitungszeit eines Testhefts konstant bleibt.

Die Erstellung eines Testdesigns läuft meist folgendermaßen ab: Als erstes legt der Testdesigner fest, welche Items in einer Studie zur Messung der Kompetenzen verwendet werden sollen und wie hoch die Personenstichprobe je Item ist. Hieraus wird unter Rückgriff auf die Itembearbeitungszeiten die insgesamt benötigte Testzeit berechnet. Teilt man diese Gesamttestzeit durch die je Testteilnehmer zur Verfügung stehende Testzeit ergibt sich die benötigte (Netto-) Personenstichprobe. Da bei der Testdurchführung die Stichprobe durch Abwesenheit (z. B. durch Krankheit) am Testtag normalerweise geringer ausfällt als anvisiert, muss je nach prognostizierter Ausfallrate die Bruttostichprobe entsprechend größer sein.

Hinsichtlich des Testdesigns müssen verschiedene Überlegungen angestellt beziehungsweise Entscheidungen getroffen werden. Ausgangspunkt ist die je Testteilnehmer zur Verfügung stehende Testzeit. Basierend auf dieser wird die Bearbeitungszeit der Blöcke und damit die Anzahl an Blöcken je Testheft festgelegt. Um die Flexibi-

lität bei der Kombination von Blöcken zu erhöhen, wird jeder Block meist nicht nur einmal, sondern mehrmals – natürlich dann in verschiedenen Testheften – verwendet. Diese Anzahl an Instanzen jedes Blocks muss geeignet festgelegt werden. Hierbei gilt, dass eine höhere Anzahl an Blockinstanzen zwar die Kombinationsmöglichkeiten erhöht, gleichzeitig jedoch auch die Anzahl an Testheften, was wiederum zu höheren administrativen Kosten führt. Die Gesamtanzahl der Blockinstanzen geteilt durch die zur Verfügung stehenden Blockpositionen je Testheft ergibt nun die Anzahl der Testhefte. Damit ist das Testdesign von den strukturellen Charakteristika her definiert: Es existiert eine Anzahl T an disjunkten Blöcken, die aus I Items gebildet wurden. Von diesen Blöcken werden jeweils R Blockinstanzen gebildet. Weiterhin sind B Testhefte mit jeweils P Positionen vorhanden. Die $T * R$ Blockinstanzen können demnach nun auf die $B * P$ Blockpositionen verteilt werden. Diese Zuweisung von Blockinstanzen zu Testheften kann durch eine Vielzahl an Anforderungen und Randbedingungen restringiert sein.

Eine wichtige Bedingung ist die Art der *Verlinkung* der Blöcke. Zwei Blöcke sind miteinander *explizit* verlinkt, wenn diese zusammen in einem Testheft auftreten. Eine *implizite* Verlinkung von zwei Blöcken liegt dann vor, wenn zwei Blöcke nicht explizit miteinander verlinkt sind, aber über andere Blöcke, die jeweils miteinander explizit verlinkt sind, verbunden sind. Eine vollständige Verlinkung eines Designs liegt dann vor, wenn jeder Block mit jedem anderen explizit oder implizit verlinkt ist. Diese Designeigenschaft ist insbesondere dann wichtig, wenn Testhefte an Teilstichproben verteilt werden, die unterschiedliche Kompetenzausprägungen aufweisen. Durch die Verlinkung wird dann nämlich sichergestellt, dass die Itemparameter nicht in Referenz zur Fähigkeit der jeweiligen Teilstichprobe an Testteilnehmern, die diese Items bearbeitet haben, geschätzt werden, sondern referenziert auf die Gesamtstichprobe.

Zwei weitere Designeigenschaften, die in Testdesigns relevant sind und deshalb auch häufig optimiert werden, sind die *Positionsbalance* und die *Blockpaarbalance*. Designs, die bezüglich einer der beiden Faktoren (Positionen oder Blockpaare) balanciert sind, werden in der Literatur *Balanced incomplete block designs* (BIBD) genannt (Frey, Hartig & Rupp, 2009; Gonzalez & Rutkowski, 2010). Zur klareren Unterscheidung solcher Designs wird vorgeschlagen, positionsbalancierte Designs *Position balanced designs* (PBD) und blockpaarbalancierte Designs *Cluster pair balanced designs* (CPBD) zu nennen. Während in PBDs jeder Block an allen Positionen gleichhäufig auftritt, kommt in CPBDs jedes Blockpaar gleich häufig vor. Obwohl *Balance* ursprünglich als dichotome Eigenschaft – ein Design ist entweder balanciert oder nicht – definiert wurde, besitzen Designs immer auch eine graduelle Balance, da sich selbst unbalancierte Designs dahingehend unterscheiden, wie stark nicht balanciert diese sind. Wie sich der Grad der Positionsbalancierung und der Blockpaarbalancierung von Designs auf die statistischen Gütemaße von Messmodellen auswirkt, ist eine offene Forschungsfrage.

1.3.6 Bias in IRT-Modellen

Ein wichtiger Kennwert zur Beurteilung eines statistischen Verfahrens ist der *Bias* eines Parameters. Dieser gibt an, wie stark der Erwartungswert des Schätzers vom wahren Wert abweicht. Für das Zustandekommen von Bias sind mehrere Gründe denkbar. Ein häufiger Grund sind Verletzungen der Voraussetzungen statistischer Verfahren. Weiterhin kann Bias aber auch auftreten, wenn wichtige Faktoren im Modell weggelassen werden (*omitted variable bias*; z. B. Greene, 2011; Wooldridge, 2013). Hierbei kompensiert das Modell für den weggelassenen Faktor dadurch, dass der Effekt anderer Faktoren über- oder unterschätzt wird. In IRT-Modellen kommt noch ein weiterer Aspekt, der zum omitted variable bias beiträgt, hinzu. In linearen

Regressionsmodellen ist die Gesamtvarianz konstant und bekannt. Durch Hinzufügen eines Prädiktors steigt die aufgeklärte Varianz, während die Fehlervarianz sinkt. Die Größe der Fehlervarianz verändert sich also in Abhängigkeit der Prädiktoren im Modell. Hingegen ist in logistischen Regressionen (und damit auch in vielen IRT-Modellen, wie zum Beispiel dem Rasch-Modell) die Fehlervarianz immer konstant, während die Gesamtvarianz modellabhängig variiert (z. B. De Boeck, 2008; Mood, 2010). Das heißt auch, dass sich die Parameter allein dadurch ändern, wenn neue Faktoren in das Modell aufgenommen werden. Bias entsteht dann auch dadurch, dass relevante Faktoren im Modell fehlen. Diese Problematik wird zum Beispiel von Mood (2010) wie auch von Weirich (in Vorbereitung) hervorragend beschrieben und illustriert.

In der vorliegenden Dissertation wird der omitted variable bias unter dem allgemeinen Begriff Bias behandelt. Insbesondere im Rahmen der Untersuchung der Positionsbalancierung von Designs wird der Bias, der durch Nichtmodellierung der Positionen zustande kommt, analysiert (Abschnitt 2.2.2 und Abschnitt 3.2.2). Auch für das Berichten von Effekten aus IRT-Modellen ergeben sich Implikationen. Da die Effekte modellabhängig unterschiedlich normiert sein können, schlagen Nakagawa und Schielzeth (2013) und Mood (2010) standardisierte Maße zur Beschreibung von Effektgrößen für IRT-Modelle vor. Diese Kennwerte werden in Abschnitt 3.1.2 verwendet, um die Größe von Positionseffekten zu berichten.

1.4 Anliegen, Ziele und Forschungsfragen

Aus den bisherigen Ausführungen in diesem Kapitel lässt sich zusammenfassend festhalten: Das Rasch-Modell wird in der Schulleistungsforschung aufgrund konzeptueller Einfachheit und Plausibilität, bestimmte Interpretationen begünstigender Eigen-

schaften und sicher auch wegen der Vielzahl an verfügbaren Softwarepaketen stark genutzt. Die Reduktion auf ausschließlich Personen- und Itemeffekte ist jedoch eine starke Vereinfachung, die in den meisten Anwendungsfällen zu kurz greift. Meist existieren eine Reihe weiterer Effekte, die im Rasch-Modell nicht mitmodelliert werden, aber trotzdem die Lösungswahrscheinlichkeit beeinflussen. Da diese Effekte im Kontext der Messung entstehen, werden diese auch unter dem Oberbegriff *Kontexteffekte* zusammengefasst. Werden solche Kontexteffekte in den statistischen Messmodellen nicht berücksichtigt, kann dies zu Verzerrungen der Modellparameter führen. Das erste Anliegen dieser Dissertation ist deshalb, relevante Kontexteffekte zu identifizieren. Die in dieser Dissertation untersuchten Kontexteffekte sind Testhefteffekte, Positionseffekte und Designeffekte, die sich alle der Gruppe der instrument effects nach Brennans Theorie (Brennan, 1992) zuordnen lassen.

Testhefteffekte kommen durch die spezifische Zusammenstellung von Testheften zustande. Das im Large-scale Assessment etablierte Vorgehen besteht in der Konstruktion von gleich schweren Testheften. Oftmals ist es allerdings wünschenswert, Testhefte mit unterschiedlichen Schwierigkeiten[3] zu konstruieren, um unterschiedlichen Personengruppen adäquatere Testhefte anbieten zu können. Hierbei besteht die Annahme, dass durch eine bessere Passung der Schwierigkeit des Testhefts und der Fähigkeit des Testteilnehmers die Motivation zur sorgfältigen Testbearbeitung erhöht wird. Durch Variation der Testheftschwierigkeiten könnten dann allerdings auch Effekte der Testhefte auftreten: Schwere Testhefte sind dann womöglich über die Itemschwierigkeiten hinaus noch zusätzlich schwieriger, während leichte Testhefte noch leichter werden. Weiterhin bestehen Testhefte meist nicht aus der gleichen Anzahl an Items, da dies aufgrund komplexer Randbedingungen beim Erstellen des

[3] Auf konzeptueller Ebene werden in dieser Dissertation die Bezeichnungen Testheftleichtigkeit und Testheftschwierigkeit gleichbedeutend verwendet. Beim Berichten von Ergebnissen wird jedoch streng auf die richtige Verwendung der Begriffe geachtet.

Testdesigns schwierig zu bewerkstelligen ist. Auch diese Variation könnte zu zusätzlichen Effekten führen, da das Bearbeiten einer größeren Anzahl an Items in der gleichen Zeit eine größere kognitive Herausforderung darstellen könnte. Die Forschungsfragen lauten deshalb: Wie viel Spielraum besteht bei der Konstruktion von Testheften? Können Testhefte mit unterschiedlicher Schwierigkeit konstruiert werden, ohne dass zusätzliche Effekte auf die Lösungswahrscheinlichkeiten auftreten? Muss die Anzahl der Items konstant gehalten werden, oder sind geringe Variationen, wie sie üblicherweise bei der Erstellung des Testdesigns entstehen, unkritisch?

Im Gegensatz zu Testhefteffekten stellen Positionseffekte ein intensiver beforschtes Themengebiet dar. Insgesamt ist gut belegt, dass Positionseffekte im Large-scale Assessment ein relevantes Phänomen sind. Das Ziel besteht deshalb auch nicht darin, die Existenz von Positionseffekten nachzuweisen. Vielmehr soll einerseits ein erweitertes Modell zur Modellierung von Positionseffekten basierend auf dem *Generalized Linear Mixed Model* (GLMM) Framework (De Boeck & Wilson, 2004) vorgeschlagen und mit diesem die Größe von Positionseffekten in den Daten des IQB-Ländervergleichs 2012 (Pant et al., 2013) bestimmt werden. Andererseits dienen diese empirisch ermittelten Positionseffekte als Grundlage für eine Simulationsstudie zur Untersuchung von Effekten der Designbalancierung.

Ähnlich wie Testhefte besitzen auch Testdesigns bestimmte Eigenschaften, die sich auf die Messung auswirken können. Die in dieser Dissertation untersuchten Designeigenschaften sind die Positionsbalance und die Blockpaarbalance. Die Untersuchung dieser beiden Eigenschaften ist sehr relevant, da quasi alle Testdesigns bestimmte Ausprägungen auf diesen Eigenschaften besitzen und die meisten großen Large-scale Assessment Programme ihre Testdesigns bezüglich dieser Eigenschaften optimieren. Wie sinnvoll ist diese Optimierung von Testdesigns bezüglich der Positions- und Blockpaarbalance? Kann hierdurch eine höhere Akkuratheit bei der

Parameterschätzung in bestimmten IRT-Modellen erreicht werden? Lapidar formuliert: Lohnt es sich, Testdesigns bezüglich Positionen und Blockpaaren zu optimieren oder kann darauf verzichtet werden?

Testhefte besitzen immer auch eine anvisierte Bearbeitungszeit, in der die Testteilnehmer die Items bearbeiten sollen. Testhefte sollten also auch in Bezug auf diese Sollbearbeitungszeit optimiert werden. Dazu müssen die zur Konstruktion von Testheften herangezogenen Item- beziehungsweise Aufgabenbearbeitungszeiten so akkurat wie möglich sein. Durch inakkurate Bearbeitungszeiten wird entweder wertvolle Testzeit nicht vollständig genutzt, wenn die Bearbeitungszeiten überschätzt wurden. Oder die Schülerinnen und Schüler schaffen die Items am Ende der Testhefte bei Unterschätzung der Aufgabenbearbeitungszeiten nicht mehr, was die Validität des Tests und die Parameterschätzungen negativ beeinflusst. Die beste, jedoch auch kostenintensivste Methode, um akkurate Aufgabenbearbeitungszeiten für das Testdesign zu erlangen, ist deren empirische Bestimmung in Vorstudien. Ist dieses Vorgehen (z. B. aus Kostengründen) nicht möglich, müssen die für das Testdesign nicht verzichtbaren Aufgabenbearbeitungszeiten anderweitig geschätzt werden. Hierzu eignet sich ein empirisch fundiertes Vorhersagemodell, in dem aus leicht verfügbaren Aufgabeneigenschaften die Bearbeitungszeiten geschätzt werden können. Das Ziel besteht deshalb in der Generierung eines solchen Vorhersagemodells, wobei folgende Forschungsfragen gestellt werden: Welche Aufgabeneigenschaften haben einen Einfluss auf die Bearbeitungszeit? Unterscheiden sich verschiedene Personengruppen in ihrer Bearbeitungszeit? Wie gut ist das generierte Vorhersagemodell?

Mit Hilfe der neuen Erkenntnisse aus dieser Dissertation können Messinstrumente im Large-scale Assessment also hinsichtlich einiger Kriterien optimiert werden. Es wird herausgearbeitet werden, ob Testhefteffekte auftreten und wodurch diese zustande kommen. Die Effektivität der Balancierung von Positionen und Blockpaaren

wird untersucht. Und es wird ein Vorhersagemodell zur Gewinnung von akkuraten Aufabenbearbeitungszeiten vorgeschlagen. Alles in allem tragen diese drei spezifischen Forschungsvorhaben dazu bei, Messungen im Large-scale Assessment präziser und vertrauenswürdiger zu machen.

2 Überblick über die Forschungsvorhaben

In diesem Kapitel werden die Forschungsvorhaben beschrieben und theoretisch eingebettet. Ebenfalls erfolgt eine Darstellung der verwendeten Methoden.

2.1 Identifikation von Kontexteffekten

2.1.1 Testhefteffekte

Die im Large-scale Assessment häufig verwendeten Multiple Matrix Sampling Designs (siehe Abschnitt 1.3.2) zeichnen sich dadurch aus, dass nicht nur ein Testheft, sondern mehrere Testhefte, die jeweils nur eine Teilmenge des gesamten Itempools beinhalten, zur Messung verwendet werden. Die Konstruktion mehrerer unterschiedlicher Testhefte impliziert aber auch, dass unterschiedliche Ausprägungen der Testhefte auf Testhefteigenschaften potentiell das Antwortverhalten beeinflussen können. Solche Testhefteffekte reihen sich daher konzeptuell in die Gruppe der Kontexteffekte ein. Obwohl Kontexteffekte vielseitig und intensiv erforscht wurden (z. B. Mollenkopf, 1950; Sax & Carr, 1962; Brenner, 1964; Sax & Cromack, 1966; Flaugher, Melton

& Myers, 1968; Munz & Smouse, 1968; Smouse & Munz, 1968; Berger, Munz, Smouse & Angelino, 1969; Smouse & Munz, 1969; Marso, 1970; Monk & Stallings, 1970; Huck & Bowers, 1972; Klosner & Gellman, 1973; Hambleton & Traub, 1974; Towle & Merrill, 1975; Whitely & Dawis, 1976; Plake, 1980; Yen, 1980; Plake, Thompson & Lowry, 1981; Plake, Ansorge, Parker & Lowry, 1982; Weiss, 1982; Kingston & Dorans, 1984; Plake, Patience & Whitney, 1988), existiert zu Testhefteffekten kaum Literatur. Lediglich in den technischen Berichten des *Programme for International Student Assessment* (PISA; OECD, 2002, 2005, 2009, 2012) werden Testhefteffekte berichtet, jedoch nicht intensiv und detailliert erörtert und diskutiert.

Die Verwendung unterschiedlicher Testhefte bietet sowohl Vor- als auch Nachteile. Ein wichtiger Vorteil besteht darin, dass Testhefte konstruiert werden können, die aus im Mittel unterschiedlich schweren Items bestehen. Diese unterschiedlich schweren Testhefte können dann unterschiedlich fähigen Teilstichproben zur Bearbeitung vorgelegt werden, so dass jede Teilstichprobe ihrer Fähigkeit entsprechende Testhefte bekommt. Dadurch steigt sowohl die Messpräzision, da diese besonders hoch ist, wenn ein Item für eine Person ein mittleres Schwierigkeitsniveau aufweist (z. B. Embretson & Reise, 2000). Weiterhin kann postuliert werden, dass motivationale und emotionale Einflüsse, wie zum Beispiel Anstrengungsbereitschaft, Langeweile und Frustration, durch eine optimale Passung zwischen Personenfähigkeit und Testheftschwierigkeit positiv beeinflusst wird. Dass solche Effekte eine Rolle spielen, wird von verschiedenen Studien nahe gelegt (Asseburg & Frey, 2013; Cole, Bergin & Whittaker, 2008; Eklöf, 2010; Wise & DeMars, 2005; Wolf & Smith, 2005). Zum Beispiel ist die gemessene Kompetenz für Schülerinnen und Schüler mit niedriger Motivation zur Testbearbeitung geringer als ihre tatsächliche Kompetenz (Wise & DeMars, 2005). Weiterhin hängt die individuelle Anstrengungsbereitschaft im Test von der Differenz der Fähigkeit und der Testheftschwierigkeit ab (Asseburg & Frey,

2013). Auf die durch inadäquate Passung potentiell verursachten Motivationseffekte gehen auch van der Linden, Veldkamp und Carlson (2004) ein: „...careless behavior is expected to increase if the items are much too easy, and motivation decreases if they are much too difficult for the students." (S. 321), und schlagen deshalb vor: „It may therefore pay off to design the assessment booklets such that each population receives items with a probability of success as close as possible to an optimum..." (S. 321). Unterstützung für diesen Vorschlag kommt aus der Forschung zu computerbasierten adaptiven Tests (*Computer adaptive testing*, CAT), deren zentrale Idee ja genau die Anpassung von Tests an Personen ist. Wise (2014) fasst die Ergebnisse der Forschung in diesem Bereich folgendermaßen zusammen: „there is evidence that the targeting of item difficulty to examinee proficiency brings with it modest motivational benefits that can improve individual score validity" (S. 13). Ein Nachteil der Verwendung verschiedener Testhefte ist jedoch, dass potentielle Effekte dieser in häufig verwendeten Messmodellen – wie zum Beispiel dem Rasch-Modell – nicht berücksichtigt werden. Dadurch gehen deren Effekte in die Schätzung der Item- und/oder Personenparameter ein und können diese verzerren. Auch Mazzeo und von Davier (2014) definieren Testhefteffekte mit Hinblick auf deren Auswirkungen: „We use the term *booklet effects* to refer to those situations in which IRT analyses do not produce the intended equivalence of results between the booklets." (S. 236). Hängen also die Ergebnisse von IRT-Modellen von den verwendeten Testheften ab, liegen Testhefteffekte vor.

Zur theoretischen Einbettung von Testhefteffekten kann Brennans Framework (Brennan, 1992) herangezogen werden (siehe auch Abschnitt 1.3.3). Laut Brennan besteht ein Instrument aus „sets of items, stimuli or tasks" (S. 237). Damit sind Testhefte Instrumente im Sinne Brennans und können somit theoretisch einen *instrument effect* verursachen. Weiterhin sind Testhefte normalerweise nicht identisch,

sondern können unterschiedlich konstruiert werden. Somit weisen diese auch unterschiedliche Ausprägungen auf verschiedenen Eigenschaften wie beispielsweise eine unterschiedliche Itemanzahl auf. Das universe of generalization besteht also aus allen möglichen Testheften, die durch Kombination aller möglichen Ausprägungen auf allen möglichen Eigenschaften definiert sind. Diese Auswahl könnte in einem universe of allowable observations eingeschränkt werden, indem zum Beispiel alle Testhefte nur eine bestimmte Itemanzahl aufweisen. Damit gibt es in beiden Universen jeweils einen Effekt des Instruments. Diese Effekte müssen nicht identisch sein. Demnach qualifizieren Testhefteffekte als Kontexteffekte im Sinne von Brennan.

Zur Modellierung von Testhefteffekten kann das im Rahmen des GLMM Frameworks dargestellte Rasch-Modell (Abschnitt 1.2.1) um einen Testheftparameter γ_b erweitert werden:

$$P(X_{jib} = 1) = \text{logit}^{-1}(\alpha_0 + \theta_j + \beta_i + \gamma_b) \tag{2.1}$$

Sowohl die Item- als auch die Testheftparameter sind in diesem Modell positiv parametrisiert. Dadurch können diese als Item- beziehungsweise Testheft*leichtigkeit* interpretiert werden.

Erfolgt die Zuweisung von Testheften zu Studienteilnehmern nicht zufällig, sondern beispielsweise in Abhängigkeit der mittleren Kompetenz von Teilstichproben (z. B. Gymnasiasten vs. Schülerinnen und Schüler anderer Schularten), sollte die Gruppenzugehörigkeit mit in das Modell aufgenommen werden. Ansonsten werden Testhefte, die nur einer Teilstichprobe zur Bearbeitung vorgelegt wurden, auch nur in Referenz zur Kompetenz dieser Teilstichprobe geschätzt, und nicht in Referenz zur Gesamtstichprobe. Das *nonequivalent groups booklet model* (NEGB) mit zusätz-

lichen Gruppenparametern δ_s lautet dann:

$$P(X_{jibs} = 1) = \text{logit}^{-1}(\alpha_0 + \theta_j + \beta_i + \gamma_b + \delta_s + \epsilon_{bs}) \tag{2.2}$$

Sollen zusätzlich Testhefteigenschaften zur Erklärung der Variation in den Testhefteffekten verwendet werden, kann das *nonequivalent groups booklet properties model* (NEGBP) verwendet werden:

$$P(X_{jibs} = 1) = \text{logit}^{-1}(\alpha_0 + \theta_j + \beta_i + \sum_{k=1}^{K} \kappa_k X_k + \nu_b + \delta_s + \epsilon_{bs}) \tag{2.3}$$

Für eine ausführliche Beschreibung dieser Modelle siehe Abschnitt *Models* im ersten Beitrag.

2.1.2 Positionseffekte

Auch für Positionseffekte lässt sich fragen, ob diese theoretisch als Kontexteffekte qualifizieren. Das universe of generalization besteht aus allen Items an allen Positionen. Dieses kann aber auch eingeschränkt sein, indem ein Test nur aus einer Teilmenge aller möglichen Itempositionierungen besteht. Dadurch kann die Itempositionierung einen Effekt auf die Lösungswahrscheinlichkeit hervorrufen. Positionseffekte sind also auch Kontexteffekte.

Dass Positionseffekte in Large-scale Assessment Studien auftreten, ist ein gut dokumentiertes Phänomen (z. B. Albano, 2013; Debeer & Janssen, 2013; Hahne, 2008; Hohensinn et al., 2008; Hohensinn, Kubinger, Reif, Schleicher & Khorramdel, 2011; Weirich, Hecht & Böhme, 2014). Das Ziel besteht deshalb auch nicht primär darin, das Auftreten von Positionseffekten zu untersuchen, sondern die Effektivität

von Strategien zur Handhabung von Positionseffekten zu evaluieren. Hierzu wird eine Simulation verwendet. Um für das Simulationsmodell realistische Positionseffekte zu gewinnen, ist es notwendig, diese vorher mit Hilfe empirischer Daten zu modellieren. Deshalb wird im Rahmen des GLMM Frameworks ein Modell zur Untersuchung von Positionseffekten vorgeschlagen und auf Daten aus dem IQB-Ländervergleich 2012 (Pant et al., 2013) angewendet. Der verwendete Datensatz ist der Teildatensatz mit $N = 19107$ Schülerinnen und Schülern, der auf dem Youden Square Design für die naturwissenschaftlichen Fächer (Hecht, Roppelt & Siegle, 2013) beruht. Das mit dem R Paket *lme4* (Bates, Mächler, Bolker & Walker, 2014b; R Core Team, 2014b) geschätzte Modell lautet:

$$P(X_{jip} = 1) = \text{logit}^{-1}(\alpha_0 + \theta_{jp} + \beta_i + \delta_p) \tag{2.4}$$

wobei $P(X_{jip} = 1)$ die Wahrscheinlichkeit einer Person j ist, das Item i an der Position p zu lösen. Der Logit dieser hängt von positionsspezifischen Personenparametern θ_{jp}, dem Itemparameter β_i, dem Positionseffekt δ_p und einem Intercept α_0 ab. Die Itemparameter werden als normalverteilt mit Mittelwert Null und Varianz σ^2_β angenommen; die Personenparameter als multivariat normalverteilt: $\theta_{jp} \sim MVN(\mathbf{0}, \mathbf{\Sigma})$ mit einem Null-Vektor $\mathbf{0}$ als Lokationen und einer Varianz-Kovarianz-Matrix:

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma^2_1 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma^2_2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma^2_p \end{pmatrix} \tag{2.5}$$

Die positionsspezifischen Varianzen σ^2_p auf der Diagonale geben an, wie stark die Personenparameter an jeder Position im Test variieren, während die Kovarianzen

$\sigma_{p_1 p_2}$ ($p_1 \neq p_2$) charakterisieren, wie stark die Personenparameter an zwei Positionen miteinander assoziiert sind.

Statt auf das häufig verwendete *Reference Coding* zur Kodierung der Positionseffekte zurückzugreifen, wird zum Zwecke einer intuitiveren Interpretierbarkeit das *Deviation Coding* (z. B. Hutcheson & Sofroniou, 1999) angewandt. Durch diese Kodierungsart kann die Abweichung je Position vom Gesamtmittelwert bestimmt werden. Hierdurch werden Aussagen wie „an Position X sind die Items um Betrag Y schwerer oder leichter als im Durchschnitt“ möglich.

2.1.3 Designeffekte

Unter einem Testdesign wird die Gesamtheit an relevanten Einheiten (Items, Testhefte, ...) und deren Beziehungen verstanden (siehe Abschnitt *Booklet Designs* im zweiten Beitrag für eine detaillierte Beschreibung und Illustration und Abschnitt 1.3.5). Im Gegensatz zu Testheften, die das Resultat des Designprozesses sind, definiert das Testdesign, nach welchen Regeln die Items zu unterschiedlichen Testheften zugewiesen werden. Testdesigns können sich also dahingehend unterscheiden, welche Regeln in welchem Ausmaß angewendet wurden. Beispielsweise können Testdesigns einen unterschiedlichen Balancierungsgrad von vollständig unbalanciert bis vollständig balanciert aufweisen. Ebenfalls kann sich die Balancierung auf verschiedene Einheiten (z. B. Itempositionen, Blockpaare) beziehen. Auch für Testdesigns gibt es also ein universe of generalization, das alle möglichen Testdesigns, die durch Kombination aller möglichen Ausprägungen auf allen möglichen Eigenschaften definiert sind, beinhaltet. Demnach lassen sich Testdesigns auch als Kontexteffekte einordnen. Da Testdesigns das Messinstrument charakterisieren, sind Kontexteffekte, die auf Eigenschaften von Testdesigns zurückzuführen sind, ebenfalls zur Gruppe der

instrument effects zuordenbar.

Verglichen mit Testheft- und Positionseffekten lassen sich Designeffekte jedoch nur schwierig empirisch untersuchen, da hierfür viele Erhebungen mit unterschiedlichen Designs nötig wären, was aus Kostengründen meist unrealistisch ist. Als Alternative bieten sich deshalb Simulationsstudien an. Die zur Untersuchung von Designeffekten verwendete Simulation ist in Abschnitt 2.2.2 beschrieben.

Zusammenfassend kann festgehalten werden, dass verschiedene Kontextfaktoren (Testhefte, Item- bzw. Blockpositionen, Testdesigns) theoretisch Einfluss auf die Messung haben können. Die in dieser Dissertation untersuchten Kontexteffekte (Testhefteffekte, Positionseffekte und Designeffekte) sind der Gruppe der *instrument effects* (Brennan, 1992) zuzuordnen. Die Identifikation dieser ist jedoch erst der erste Schritt. Zur Optimierung von Messinstrumenten ist nicht nur wichtig zu wissen, dass Kontexteffekte auftreten, sondern auch warum. Es muss also eruiert werden, welche Eigenschaften der Kontextfaktoren für die Kontexteffekte verantwortlich sind.

2.2 Optimierung von Messinstrumenten

Zusätzlich zur Identifikation von Faktoren, die Kontexteffekte im Rahmen der Messung erzeugen, ist es wichtig zu ergründen, auf welche Eigenschaften der Faktoren die Effekte zurückzuführen sind. Dadurch lassen sich Erkenntnisse zur Optimierung von Messinstrumenten ableiten. Die in dieser Dissertation untersuchten Eigenschaften sind die Schwierigkeit und die Itemanzahl von Testheften und der Balancierungsgrad von Testdesigns bezüglich Positionen und Blockpaaren. Inakkurate Aufgabenbearbeitungszeiten können ebenfalls zu Kontexteffekten durch Zeitmangel führen. Deshalb ist es wichtig, möglichst akkurate Bearbeitungszeiten für das Testdesign

zur Verfügung zu stellen. Dies kann durch ein empirisch fundiertes Vorhersagemodell für Aufgabenzeiten geleistet werden.

2.2.1 Testheftschwierigkeit und Itemanzahl

Neben der Untersuchung, ob Testhefteffekte auftreten, soll ebenfalls der Frage nachgegangen werden, durch welche Eigenschaften der Testhefte diese Effekte erzeugt werden. Hieraus lassen sich Implikationen für die Optimierung von Messinstrumenten im Large-scale Assessment ableiten. Testhefteigenschaften, die keinen Effekt erzeugen, brauchen im Testdesign oder im Messmodell nicht berücksichtigt werden. Hingegen sollte Eigenschaften, die einen Effekt auf das Antwortverhalten ausüben, im Designprozess mittels der in Abschnitt 1.3.4 beschriebenen Methoden begegnet werden. Zur Untersuchung des Effekts von Testhefteigenschaften müssen diese jeweils variieren, da ansonsten deren Effekt nicht modellierbar ist (Weirich, Hecht & Böhme, 2014). Die beiden Testhefteigenschaften, die untersucht werden sollen, sind die *A-priori-Schwierigkeit* und die Anzahl an Items des Testhefts. Die A-priori-Schwierigkeit ist die vor der Erhebung geschätzte Schwierigkeit des Testhefts. Diese Schätzung kann beispielsweise durch die Mittelung von Itemschwierigkeiten aus Vorstudien erfolgen. Die Annahme ist, dass Testhefte, die im Design vorab als schwieriger konzipiert wurden, dies auch tatsächlich sind. Hierbei muss betont werden, dass es sich bei der empirisch untersuchten Testheftschwierigkeit immer um die für Itemschwierigkeiten kontrollierte, „zusätzliche“ Testheftschwierigkeit handelt. Die Anzahl an Items könnte ebenfalls dazu führen, dass unterschiedliche Testhefte unterschiedlich schwer sind. Testhefte mit vielen Items könnten hierbei überdurchschnittlich schwer sein, da höhere kognitive Ressourcen erforderlich sind.

Bezüglich der Optimierung von Messinstrumenten können sich folgende Szena-

rien ergeben: Zeigen sich für die beiden untersuchten Testhefteigenschaften Nulleffekte, können diese beliebig variiert werden. Dies hätte den Vorteil, dass man bei der Erstellung des Designs mehr Freiheiten hätte. Einerseits müsste nicht mehr darauf geachtet werden, wie viele Items einem Testheft zugewiesen werden. Andererseits könnte die Testheftschwierigkeit intendiert variiert werden, um bestimmte Personengruppen mit adäquateren Testheften, die von ihrer Schwierigkeit im Mittel besser zu den Personenfähigkeiten passen, zu versorgen. Sollte sich hingegen zeigen, dass die beiden Testhefteigenschaften einen Einfluss auf die Lösungswahrscheinlichkeiten ausüben, können mehrere Strategien verfolgt werden.

Eine Strategie wäre, den Test zu standardisieren, indem man die Itemanzahl und die Testheftschwierigkeit konstant hält. Dies schränkt jedoch die Generalisierbarkeit dahingehend ein, dass die Ergebnisse nicht mehr auf Tests mit anderen Ausprägungen auf diesen Testhefteigenschaften generalisiert werden können. Soll für eine weitere Large-scale Assessment Studie wieder ein Test aus dem gleichen Itempool zusammengestellt werden, muss darauf geachtet werden, dass die Itemanzahl und die Testheftschwierigkeit jenen aus den Vorstudien entspricht.

Ist Konstanthaltung der Testhefteigenschaften nicht möglich oder erwünscht, können deren Effekte auch durch Balancierung oder Randomisierung „ausgemittelt“ werden. Hierbei muss allerdings die Einheit, über die Aussagen gemacht werden sollen, mit in Betracht gezogen werden. Soll die Fähigkeit einer bestimmten Person untersucht werden, würde Balancierung bedeuten, dass diese Person Testhefte mit allen möglichen Ausprägungen auf den Testhefteigenschaften bearbeitet, zum Beispiel also Testhefte mit allen möglichen Anzahlen an Items. Auch bei Randomisierung, also zufälliger Ziehung von Testheften aus der Population von Testheften, müsste eine Person dann mehrere bis viele Testhefte bearbeiten. Dies ist aus testadministrativen Gründen quasi ausgeschlossen. Die Analyseeinheit in Large-scale Assessments sind

jedoch auch nicht Einzelpersonen, sondern Personengruppen, deren mittlere Fähigkeit von Interesse ist. Deshalb bietet sich hier eine Randomisierung an, indem die Testhefte den Testteilnehmern zufällig zugeordnet werden. Dadurch mitteln sich Effekte der Testhefteigenschaften bei der Bestimmung der mittleren Personenfähigkeit heraus. Aber auch hierbei stellt sich die Frage der Generalisierbarkeit, da die eingesetzten Testhefte normalerweise nicht zufällig aus der Population aller Testhefte gezogen werden, sondern aus einer bereits eingeschränkten Teilmenge der Population. Zum Beispiel bewegt sich die Testheftanzahl meist in engeren Grenzen als möglich wäre, da Testhefte mit nur wenigen Items aus erhebungstechnischen Gründen selten eingesetzt werden. Deshalb kann auch nur auf die Teilpopulation an Testheften mit den normalerweise üblichen Eigenschaften generalisiert werden.

2.2.2 Balancierung von Positionen und Blockpaaren

Die in Testdesigns im Large-scale Assessment oft balancierten Faktoren sind die Positionen der Blöcke in den Testheften und das Auftreten von Blockpaaren. In den folgenden Abschnitten wird deren Relevanz und die Methode zu deren Untersuchung skizziert.

Die Relevanz von Positionseffekten und die Notwendigkeit ihrer Balancierung wird in zwei Anwendungsfällen besonders deutlich. Zwei Items mit derselben Schwierigkeit aber unterschiedlicher Positionierung werden in Messmodellen, in denen keine Positionseffekte mitmodelliert werden (z. B. im Rasch-Modell), auf unterschiedliche Schwierigkeiten geschätzt. Für Methoden, die auf eine korrekte Itemsortierung nach Schwierigkeiten angewiesen sind – wie beispielsweise die *Bookmark Method* (Mitzel, Lewis, Patz & Green, 2001) in *Standard Setting* Verfahren (z. B. Cizek & Bunch, 2007) – ist die Verzerrung von Itemparametern durch Positionseffekte sehr nachtei-

lig. Ein weiterer Anwendungsfall, der negativ durch verzerrte Itemparameter beeinflusst wird, sind *Linking* Verfahren mit *Common Items* (z. B. Dorans, Pommerich & Holland, 2007). Wird dasselbe Item in zwei Studien an verschiedenen Positionen platziert, gefährden daraus resultierende Unterschiede in der Itemschwierigkeit das Linking (Meyers, Miller & Way, 2009). Obwohl Konsens besteht, dass Positionseffekte ein relevantes Phänomen sind und die Vermeidung negativer Effekte durch deren Balancierung möglich ist, wurde bisher wenig erforscht, wie stark sich ein unterschiedlicher Balancierungsgrad des Designs auf die Parameterschätzung auswirkt. Die Annahme ist, dass mit höherer Positionsbalancierung die Akkuratheit der Itemparameter im Rasch-Modell steigt.

Den Effekten der Blockpaarbalance wurde im Kontext und in der Terminologie der Forschung zu Testdesigns im Large-scale Assessment bisher wenig Aufmerksamkeit zuteil. Dies zeigt sich einerseits im Mangel an Literatur zur Modellierung von Blockpaareffekten in empirischen Daten. Andererseits fehlen auch in der Literatur zu Testdesigns Begründungen, warum Blockpaare balanciert werden sollten. Nichtsdestotrotz kommen vollständig oder partiell blockpaarbalancierte Designs in vielen Large-scale Assessment Studien zum Einsatz, zum Beispiel in PISA (OECD, 2012), in NAEP (Allen, Donoghue & Schoeps, 2001) und in den IQB-Studien zu den deutschen Bildungsstandards (Hecht, Roppelt & Siegle, 2013; Weirich, Haag & Roppelt, 2012; Böhme et al., 2010). Deshalb besitzt die Erforschung der Blockpaarbalancierung eine hohe Relevanz.

Die Blockpaarbalance könnte aber im Prinzip auch als *Missing data* Problem aufgefasst werden. In Abhängigkeit davon, wie viele Blockpaare in einem Design realisiert werden – also wie hoch die Blockpaarbalancierung ist –, steigt oder sinkt die verfügbare Itemkovarianzinformation. Die in populären IRT-Softwares (z. B. ConQuest, Wu, Adams, Wilson & Haldane, 2007; TAM, Kiefer, Robitzsch & Wu, 2014)

eingesetzten EM Algorithmen (Dempster, Laird & Rubin, 1977; Bock & Aitkin, 1981) können solche designbedingten Missings im Normalfall eigentlich problemlos bewältigen. Dennoch ist offen, wie stark die Blockpaarbalance und damit die Itemkovarianzinformation reduziert werden kann, bevor Probleme auftreten.

Zur Erforschung des Einflusses der Positions- und Blockpaarbalance auf die Schätzung von Itemparametern werden 1540 verschiedene Designs basierend auf einem Youden Square Design mit 31 Testheften, 31 Blöcken und 6 Positionen (siehe Table 1 im zweiten Beitrag) mittels eigens für diesen Zweck programmierten Optimierungsalgorithmen generiert (siehe Abschnitt *Generating Designs* im zweiten Beitrag).

Zur Operationalisierung der Blockpaarbalance wird der relative Anteil von realisierten an allen potentiell möglichen Blockpaaren multipliziert mit 100 verwendet. Die Blockpaarbalance rangiert somit im Bereich von 0 (unbalanciert) bis 100 (balanciert). Zur Berechnung der Positionsbalance wird die Korrelation von Positionen und Blöcken in der Design-Matrix herangezogen (siehe Frey, Hartig & Rupp, 2009). Wenn Positionen und Blöcke komplett gekreuzt sind – wenn also jeder Block an jeder Position auftritt – ist diese Korrelation 0. Wenn jede Position nur von einem Block besetzt ist und jeder Block nur an einer Position platziert wird, ist diese Korrelation gleich 1. Diese Korrelation charakterisiert demnach die *Un*balanciertheit des Designs. Um die Polarität dieser Eigenschaft zu invertieren und äquivalent zur Blockpaarbalance zu skalieren, wurde diese Korrelation so transformiert, dass die Positionsbalance-Skala ebenfalls von 0 (unbalanciert) bis 100 (balanciert) verläuft. Mit dem R-Paket *eatDesign* (Hecht, 2014) lassen sich diese beiden Kennwerte sehr einfach berechnen.

In den erzeugten Designs verläuft die Blockpaarbalance von 32 bis 100 in Schritten von 2; die Positionsbalance von 14 bis 100 ebenfalls in Schritten von 2 (siehe

Figure 2 im zweiten Beitrag). Eine höhere Unbalanciertheit der Designs ist in diesem Setting nicht möglich. Da die Testhefte mehr als einen Block enthalten, werden auch immer automatisch bestimmte Blockpaare generiert, wodurch die Blockpaarbalance nicht auf 0 sinken kann. Die Positionsbalancierung wird dadurch beschränkt, dass bei 6 Positionen und 31 Blöcken einige Blöcke immer an mehr als einer Position platziert werden müssen, da 31 kein ganzzahliges Vielfaches von 6 ist. Dadurch wird eine Balancierung größer als 0 erzeugt.

Während die beiden Balancierungen der Designs die unabhängigen Variablen sind, werden als abhängige Variablen Kennwerte herangezogen, die die Güte der Parameterschätzung kennzeichnen. Der Bias gibt an, wie stark eine Parameterschätzung vom wahren Wert im Mittel abweicht. Solche Abweichungen können verschiedene Ursachen haben (siehe Abschnitt 1.3.6). In dieser Dissertation wird der Bias, der durch Weglassen von relevanten Faktoren im Modell entsteht (omitted variable bias) unter dem allgemeinen Begriff Bias subsumiert. Während der Bias die über mehrere Replikationen gemittelte Abweichung der geschätzten Parameterwerte vom wahren Wert kennzeichnet, beinhaltet der *Root Mean Square Error* (RMSE) zusätzlich die Variabilität der Schätzungen. Wenn die Schätzung unverzerrt ist (d. h. wenn der Bias gleich null ist), charakterisiert der RMSE die Varianz, die durch das Sampling entsteht. Im Fall eines verzerrten Parameterschätzers (Bias größer null) kombiniert der RMSE die Verzerrung und die Dispersion in einem Gesamtmaß.

Die beiden Kennwerte Bias und RMSE charakterisieren also die Akkuratheit der Schätzung eines einzelnen Parameters. Um jedoch Aussagen über Testdesigns, die eine Vielzahl an Parametern enthalten, treffen zu können, müssen geeignete aggregierte Kennwerte verwendet werden. Hierzu bietet sich eine Mittelung der absoluten Verzerrungen und der RMSE an (siehe Formeln im Abschnitt *Model and Outcomes* im zweiten Beitrag). Diese Kennwerte geben nun an, wie hoch die Akkuratheit der

Parameterschätzungen in einem Design im Mittel ist. Im nächsten Schritt können diese Akkuratheitskennwerte auf Design-Ebene dann mit den Designeigenschaften Blockpaarbalance und Positionsbalance in Beziehung gesetzt werden. Somit können Erkenntnisse darüber gewonnen werden, ob die Güte der Parameterschätzung vom Ausmaß der Balancierung des Designs abhängt. Wenn die Schätzgüte von der Balancierung abhängt und das am stärksten balancierte Design auch die höchste Schätzgüte aufweist, kann empfohlen werden, bei der Zusammenstellung zukünftiger Designs besonders auf diese Balancierung zu achten. Ist die Güte der Parameterschätzung allerdings unabhängig vom Balancierungsgrad, kann die Balancierung vernachlässigt werden.

2.2.3 Vorhersagemodell für Aufgabenbearbeitungszeiten

Der Ausgangspunkt bei der Konstruktion von Messinstrumenten im Large-scale Assessment ist die Zeit, die je Testteilnehmer und Testteilnehmerin zur Bearbeitung des Tests zur Verfügung steht. Items und Aufgaben müssen so zusammengestellt werden, dass diese Sollbearbeitungszeit des Testhefts so gut wie möglich erreicht wird. Deshalb ist es von großer Wichtigkeit, die Bearbeitungszeiten der Items beziehungsweise Aufgaben möglichst genau zu kennen. Hierzu gibt es mehrere Möglichkeiten. Die akkuratesten Bearbeitungszeiten können sicherlich aus direkten Messungen gewonnen werden. Diese Methode ist allerdings auch aufwändig und kostenintensiv. Deshalb werden Bearbeitungszeiten stattdessen häufig von Experten (zum Beispiel von Aufgabenentwicklern) geschätzt. Den Erfahrungen aus dem Projekt „Evaluation der Bildungsstandards in den Naturwissenschaften für die Sekundarstufe I“ am Institut zur Qualitätsentwicklung im Bildungswesen (IQB) zufolge sind diese Schätzungen allerdings nicht immer akkurat und teilweise inkonsistent. Eine vielversprechende Alternative ist, die Bearbeitungszeiten aus zur Verfügung stehenden Daten – näm-

lich bestimmten Aufgabeneigenschaften wie zum Beispiel die Anzahl an Wörtern – zu schätzen.

Obwohl zum Thema Bearbeitungszeiten eine große Menge an Forschungsarbeiten existiert (siehe das Buchkapitel von Schnipke & Scrams, 2002; und den Zeitschriftenbeitrag von Lee & Chen, 2011), wurde die Idee der Schätzung von Bearbeitungszeiten aus Aufgaben- oder Itemeigenschaften eher selten verfolgt. Halkitis, Jones und Pradhan (1996) untersuchten den Zusammenhang zwischen Itembearbeitungszeit und Itemschwierigkeit, Itemdiskrimination und Wortanzahl. Hierbei zeigte sich, dass diese Prädiktoren die Hälfte der Varianz der Bearbeitungszeit erklärten, wobei Wortanzahl der stärkste Prädiktor ($R^2 = 27.2\%$) war, gefolgt von Itemschwierigkeit ($R^2 = 16.2\%$) und Itemdiskrimination ($R^2 = 6.8\%$). Bergstrom, Gershon und Lunz (1994) identifizierten Textlänge, (relative) Itemschwierigkeit, Itemreihenfolge und Position der richtigen Antwort (in Multiple-Choice-Items) als relevant. In einer Studie von Swanson, Case, Ripkey, Clauser & Holtman (2001) konnten 45% der Itembearbeitungszeit auf die Schwierigkeit, das Vorhandensein einer Abbildung und die Wortanzahl zurückgeführt werden. Die Autoren geben an, dass „a logit change in item difficulty adds 14+ seconds“, „the presence of a picture adds 12+ seconds“, und „each word adds approximately 0.5 seconds“ (S. 116). Insgesamt legen diese Befunde nahe, dass die Vorhersage von Bearbeitungszeiten durch Eigenschaften der Aufgaben oder Items ein aussichtsreicher Ansatz ist.

Testinstrumente im Large-scale Assessment werden immer für den Einsatz in einer spezifischen Personenstichprobe konstruiert. Deshalb ist es von Vorteil, über Informationen zum Zeitbedarf verschiedener Personengruppen zu verfügen, um Tests für diese anzupassen. In der Literatur zeigt sich insgesamt jedoch ein uneinheitliches Bild zum Zusammenhang von Bearbeitungszeit und Personeneigenschaften (siehe auch Schnipke & Scrams, 2002; Lee & Chen, 2011). Es ist demnach nicht hinreichend

geklärt, wie der Zeitbedarf verschiedener Personengruppen aussieht.

Angelehnt an vorherigen Arbeiten soll eine empirisch fundierte, leicht zu benutzende Formel zur Berechnung von Bearbeitungszeiten erstellt werden, die zur Konstruktion von Testheften mit akkuraten Sollbearbeitungszeiten herangezogen werden kann. Konkret werden die empirischen Bearbeitungszeiten durch die Aufgabeneingenschaften Anzahl an Items, Anzahl an Wörtern, Antwortformat (Anzahl an Items mit Antwortformat Multiple-Choice, Kurzantwort und Erweiterte Antwort) und Aufgabenschwierigkeit und den Personeneigenschaften Geschlecht, Schulart und Kompetenz vorhergesagt. Dieses Modell wird mithilfe einer weiteren Stichprobe unter Verwendung neuer Aufgaben validiert.

Analog zum, zur Modellierung von Testhefteffekten und Positionseffekten verwendeten, RPRI Rasch-Modell werden im ersten Modell die Einflüsse von Personen und Aufgaben entflochten. Die abhängige Variable ist jetzt allerdings nicht die richtige oder falsche Beantwortung, sondern die Bearbeitungszeit. Statt dichotomen Daten liegen also kontinuierliche Daten vor, womit man sich nicht im GLMM, sondern im *Linear Mixed Model* (LMM) Framework (z. B. Bates, 2010) bewegt. Die Idee der Modellierung unterschiedlicher Varianzquellen bleibt aber bestehen. Die abhängige Variable, die Bearbeitungszeit y_{jt} einer Person j, die eine Aufgabe t bearbeitet, wird in einen Personenparameter θ_j und einen Aufgabenparameter β_j zerlegt:

$$y_{jt} = \alpha_0 + \theta_j + \beta_t + \epsilon_{jt} \tag{2.6}$$

In diesem Modell stellt das Intercept α_0 den Gesamtmittelwert der Bearbeitungszeiten, θ_j die Abweichung der Person j von diesem Mittelwert, β_t die Abweichung der Aufgabe t und ϵ_{jt} die Interaktion der Person j mit der Aufgabe t dar. Da das Ziel in der Untersuchung von Aufgaben- und Personeneigenschaften liegt, sind die einzel-

nen Personen- und Aufgabenparameter von geringerem Interesse. Deshalb werden diese als *random effects* mit Mittelwerten von 0 und den Varianzen σ^2_θ, σ^2_β und σ^2_ϵ modelliert. Diese Varianzen geben an, wie stark die jeweiligen Einheiten bezüglich der Bearbeitungszeit variieren. Als nächstes können diese Varianzen durch Hinzunahme von Personen- und Aufgabenprädiktoren und deren Interaktionen erklärt und die Varianzaufklärung der Prädiktoren quantifiziert werden. Das finale Modell beinhaltet alle relevanten (d. h. signifikanten oder bedeutsamen) Aufgaben- und Personeneigenschaften und alle signifikanten Aufgaben × Personen-Interaktionen.

Die Stichprobe bestand aus $N = 334$ Schülerinnen und Schülern der 9. Klasse (Alter: $M = 15.5$ ($SD = 0.75$) Jahre, 49.4% Mädchen) aus vier Gymnasien (56.9%) und drei Regelschulen (43.1%). Diesen wurde zufällig eins von 24 Testheften, die insgesamt 93 Aufgaben aus dem Bereich Naturwissenschaften enthielten, zugeteilt. Die Messung der Bearbeitungszeiten erfolgte mittels Eintragung der aktuellen Uhrzeit vor und nach jeder Aufgabe durch die Testteilnehmer (siehe auch Abschnitt *Design and Procedure* auf S. 132–133). Eine zweite Stichprobe von $N = 1386$ Schülerinnen und Schülern bearbeitete zu einem späteren Zeitpunkt $N = 125$ neue Naturwissenschaftsaufgaben. Mit Hilfe der aus den Daten der ersten Stichprobe abgeleiteten Formel wurden die Aufgabenzeiten für die neuen Aufgaben vorhergesagt und mit den empirisch ermittelten Aufgabenzeiten aus der zweiten Stichprobe zu Validierungszwecken verglichen.

3 Ergebnisse

3.1 Kontexteffekte

3.1.1 Testhefteffekte

Das Ziel der Untersuchung von Testhefteffekten bestand zum einen darin, deren Relevanz im Large-scale Assessment zu eruieren. Zum anderen sollten neue Erkenntnisse zur Optimierung von Messinstrumenten gewonnen werden. Zu diesem Zweck wurde das Rasch-Modell um Testhefteffekte erweitert. Die Ergebnisse des NEGB Modells sind im ersten Beitrag dargestellt. Der im Fokus stehende Effekt ist die Schätzung der Standardabweichung der Testheftparameter. Diese beträgt $SD = 0.08$ mit einem 95%-Konfidenzintervall mit der unteren Grenze $LL = 0.05$ und der oberen Grenze $UL = 0.12$. Testhefte variieren demnach tatsächlich in ihrer Schwierigkeit (unter Kontrolle der Itemschwierigkeiten). Um die Größe dieses Effektes zu verdeutlichen, kann unter Rückgriff auf die Modellparameter der Anstieg der Lösungswahrscheinlichkeit eines Items bestimmt werden, dass vom schwierigsten in das leichteste Testhefte verschoben wird (siehe Abschnitt *Results* im ersten Beitrag). Dieser Anstieg beträgt 5.3%, eine Größenordnung, die durchaus als substantiell bewertet werden kann.

3.1.2 Positionseffekte

Table 3 im zweiten Beitrag enthält die Ergebnisse des Modells zur Modellierung von Positionseffekten (*Position effects model*). Die Positionseffekte liegen im erwarteten Größenbereich und sinken erwartungsgemäß im Verlauf des Tests. Das heißt, dass die Items weniger leicht – also schwieriger – mit zunehmender Position im Test werden. Nach der Hälfte des Tests (d. h. nach der 3. Position) wurde eine 15-minütige Pause bei der Testbearbeitung eingelegt. Hierdurch konnten sich die Schülerinnen und Schüler etwas erholen. Dies schlägt sich in den Positionseffekten insofern nieder, als dass Items an Position 4 wieder etwas leichter sind als an Position 3. Die zunehmende Schwierigkeit tritt aber ebenfalls auch wieder von Position 4 bis 6 ein. Bezüglich der positionsspezifischen Streuungen zeigt sich, dass diese im Verlauf des Tests zunehmen.

Zur besseren Interpretierbarkeit und Vergleichbarkeit wurden die Positionseffekte standardisiert, deren Varianzerklärung $R^2_{GLMM(c)}$ (Nakagawa und Schielzeth, 2013) berechnet und im Effektstärkemaß f^2 (Cohen, 1988) ausgedrückt. Eine Standardisierung der Effekte ist in IRT-Modellen besonders wichtig, da die Effekte an einer konstanten Fehlervarianz normiert werden. Dadurch variiert die Effektgröße in Modellen mit unterschiedlicher Aufklärungsstärke (siehe auch Abschnitt 1.3.6). Die standardisierten Positionseffekte (in der Parametrisierung als „Leichtigkeit“) befinden sich im Bereich von 0.09 Logits an der ersten Position bis -0.08 Logits an der sechsten Position. Die Varianzaufklärung durch die Positionseffekte beträgt 1.8%, was Cohen's $f^2 = 0.033$ entspricht. Diese Effektstärke fällt in den Bereich eines kleinen Effekts ($0.02 \leq f^2 < 0.15$).

3.1.3 Designeffekte

Zur Untersuchung von Designeffekten wurden 1540 Designs, die in den Designeigenschaften Positionsbalance und Blockpaarbalance variieren, untersucht. Die abhängige Variable war die Schätzgüte der Itemparameter im Rasch-Modell. Diese Schätzgüte wurde durch zwei Kennwerte operationalisiert: dem mittleren (absoluten) Bias der Itemparameter ($bias_\beta$) und dem mittleren RMSE der Itemparameter ($RMSE_\beta$). Zwei Datenbedingungen kamen in der Simulation zum Einsatz: (1) Rasch-konforme Daten und (2) Daten mit Positionseffekten. Die deskriptiven Ergebnisse der Schätzgütekriterien in Abhängigkeit der Datenbedingung sind in Table 4 im zweiten Beitrag dargestellt. In den Rasch-Daten ist die Verzerrung der Itemparameterschätzung quasi null ($bias_\beta = 0.013$). In den simulierten Daten mit Positionseffekten beläuft sich die Verzerrung hingegen auf $bias_\beta = 0.074$ und ist somit gut 5-mal größer. Bezüglich der Streuung der Verzerrungen der Designs zeigt sich ein ähnliches Bild. In den Rasch-Daten ist die Streuung der Verzerrungen so gut wie null ($SD_{bias_\beta} = 0.001$). Das Rasch-Modell ist demnach erwartungsgemäß extrem adäquat bei der Modellierung von Rasch-konformen Daten unabhängig vom verwendeten Design. Dagegen zeigt sich in den Positionseffekte-Daten eine Streuung der Verzerrung von $SD_{bias_\beta} = 0.029$. Es existieren also Designs, in denen die Schätzung der Itemparameter im Rasch-Modell weniger verzerrt ist als in anderen Designs. Um diese Designs zu identifizieren wurden im nächsten Schritt die beiden potentiell relevanten Designeigenschaften der Positions- und Blockpaarbalance mit den Schätzgütekriterien in Verbindung gebracht (Abschnitt 3.2.2).

3.2 Zur Optimierung von Messinstrumenten verwendbare Ergebnisse

3.2.1 Testheftschwierigkeit und Itemanzahl

Während das NEGB Modell (siehe Abschnitt 3.1.1) klar gezeigt hat, dass Testhefte in ihrer um die Schwierigkeit der Items bereinigten Schwierigkeit variieren, wurde im NEGBP Modell untersucht, worauf diese Varianz zurückzuführen ist. Hierzu wurden in diesem Modell die beiden Testhefteigenschaften A-priori-Leichtigkeit und die Anzahl an Items hinzugefügt. Da in der verwendeten Software, dem R Paket *lme4* (Bates, Mächler, Bolker & Walker, 2014a; R Core Team, 2014a) für die Parameter lediglich eine „+"-Parametrisierung angeboten wird, ändert sich die Interpretation der Parameter von *Schwierigkeit* zu *Leichtigkeit*. Die im NEGBP Modell geschätzten Effekte der beiden Testhefteigenschaften betragen $b = 0.14$, 95% CI [0.08, 0.21], für die A-priori-Leichtigkeit und $b = -0.014$ $[-0.020, -0.011]$ für die Anzahl an Items. Die Variation dieser beiden Eigenschaften im Testdesign hat also Wirkung gezeigt. Erhöht man die mittlere Leichtigkeit der Items eines Testhefts um 1 Logit, steigt die tatsächliche Leichtigkeit des Testhefts um 0.14 Logits. Das heißt, dass Testhefte aus leichten Items noch *zusätzlich* leichter und Testhefte, die schwierige Items enthalten, *zusätzlich* schwerer sind. Die spezifische Zusammenstellung eines Testhefts aus unterschiedlich schwierigen Items macht demnach tatsächlich einen Unterschied aus. Weiterhin steigt die Testheftschwierigkeit um 0.014 Logits je zusätzlichem Item. Ein Item, das im kürzesten Testheft ($N = 42$ Items) der analysierten Daten platziert wird, wäre damit $0.014 * 22 = 0.31$ Logits leichter als im längsten Testheft ($N = 64$ Items). Zusammenfassend kann festgehalten werden, dass kleine, aber substantielle Testhefteffekte existieren und dass die Testhefteigenschaften A-priori-Leichtigkeit

und Itemanzahl mit diesen Effekten in Zusammenhang stehen. Damit ergibt sich die Möglichkeit beziehungsweise Notwendigkeit, Testhefteffekte durch verschiedene Strategien zu handhaben. Diese werden im Diskussionsteil in Abschnitt 4.2 dargestellt.

3.2.2 Balancierung von Positionen und Blockpaaren

Die Schätzgüte von Itemparametern im Rasch-Modell variiert bezüglich des verwendeten Designs (siehe Abschnitt 3.1.3). Die 1540 untersuchten Designs unterscheiden sich bezüglich ihrer Positions- und Blockpaarbalance. Um zu untersuchen, ob diese Designeigenschaften mit den unterschiedlichen Schätzgüten zusammenhängen, wurde in einer multiplen Regression die (standardisierte) Variable $bias_\beta$ durch die Positionsbalance und die Blockpaarbalance vorhergesagt (Table 5 im zweiten Beitrag. Hierbei zeigt sich, dass die Verzerrungen fast komplett auf die Designbalance zurückzuführen sind ($R^2 = 95.1\%$). Der Regressionskoeffizient der Positionsbalance $b = -0.036$ $(p < .001)$ gibt an, dass die Verzerrung um -0.036 Standardabweichungen sinkt, wenn die Positionsbalance um 1 steigt. In anderen Worten, pro Punkt auf der von 0 bis 100 verlaufenden Positionsbalance-Skala verringert sich die Verzerrung um im Mittel $b * SD_{bias_\beta} = 0.036 * 0.029 = 0.001$. Zusätzlich ist die Verzerrung in den vollständig balancierten Designs mit $M_{bias_\beta} = 0.02$ wesentlich geringer als in den $pb = 14$ Designs ($M_{bias_\beta} = 0.12$). Insgesamt kann also geschlussfolgert werden, dass die Positionsbalancierung von Designs zu einer Reduktion der Verzerrungen bei der Parameterschätzung führt. Bezüglich der Blockpaarbalancierung zeigte sich hingegen ein Nulleffekt. Als Empfehlung zur Erstellung von Designs im Large-scale Assessment lässt sich demnach festhalten, dass Designs eine hohe Positionsbalance aufweisen sollten, wohingegen die Blockpaarbalance vernachlässigbar ist. Die einschränkenden Bedingungen zu dieser Aussage werden im Diskussionsteil

in Abschnitt 4.3 erläutert.

3.2.3 Vorhersagemodell für Aufgabenbearbeitungszeiten

Als kostengünstige Alternative zu aufwändigeren Verfahren, wie zum Beispiel Pilotstudien, wurde auf empirischer Datengrundlage eine Formel zur Berechnung von Bearbeitungszeiten vorgeschlagen, die zur Konstruktion von Testinstrumenten im Large-scale Assessment herangezogen werden kann. Die Bearbeitungszeiten werden hierbei von leicht verfügbaren Aufgabenmerkmalen vorhergesagt. Weiterhin wurde untersucht, inwiefern spezifische Personengruppen einen differentiellen Zeitbedarf haben, der dann bei der Zusammenstellung von Tests für diese berücksichtigt werden kann. Das finale Modell, das 94.3% der aufgabenbedingten Bearbeitungszeit erklärt, lautet:

$$\begin{aligned}\hat{y} = 43.8N_{items} + 0.39N_{words} - 25.9N_{MC} + 2.2N_{SR} + 14.7N_{ER} \\ +6.1Z_{sex}N_{ER} + 16.6Z_{track}N_{ER}\end{aligned} \tag{3.1}$$

wobei N_{items} die Anzahl an Items in der Aufgabe, N_{words} die Anzahl der Wörter der Aufgabe, N_{MC} die Anzahl an Multiple-Choice-Items, N_{SR} die Anzahl an Kurzantwort-Items, N_{ER} die Anzahl an Erweiterte-Antwort-Items, Z_{sex} das effektkodierte Geschlecht (-1 für weiblich; 1 für männlich) und Z_{track} die effektkodierte Schulart (-1 für nicht-gymnasiale Schularten; 1 für Gymnasium) ist.

Diese Formel kann verwendet werden, um Bearbeitungszeiten für (a) Stimuli, (b) Items und (c) Aufgaben zu schätzen. Im einfachsten Fall eines Stimulus sind alle Prädiktoren gleich null – außer der Anzahl an Wörtern. Zum Beispiel beträgt die Bearbeitungszeit für einen Stimulus mit 100 Wörtern $\hat{y}_{(a)} = 0.39 * 100 = 39$ Sekunden. Ein Item mit dem Antwortformat *Erweiterte Antwort* mit 100 Wörtern benötigt

$\hat{y}_{(b)} = 43.8 * 1 + 0.39 * 100 + 14.7 * 1 = 97.5$ Sekunden. Wenn dieses Item in Gymnasien eingesetzt werden soll, können 16.6 Sekunden addiert werden: $\hat{y}_{(b)2} = \hat{y}_{(b)} + 16.6 * Z_{track2} * N_{ER} = 97.5 + 16.6 * 1 * 1 = 114.1$ Sekunden. Für den Einsatz in nicht-gymnasialen Schularten werden 16.6 Sekunden substrahiert: $\hat{y}_{(b)1} = \hat{y}_{(b)} + 16.6 * Z_{track1} * N_{ER} = 97.5 + 16.6 * (-1) * 1 = 80.9$ Sekunden. Für eine Aufgabe kann auf zwei verschiedenen Wegen die Bearbeitungszeit geschätzt werden; entweder durch Anwendung der Formel oder durch Summation der berechneten Bearbeitungszeiten der einzelnen Elemente. Unter Verwendung der eben verwendeten Elemente wird eine Aufgabe mit einem Stimulus und zwei Items konstruiert. Deren Bearbeitungszeit beträgt $\hat{y}_{(c)1} = 43.8 * 2 + 0.39 * (100 + 100 + 100) + 14.7 * 2 = 234$ Sekunden. Alternativ können aber auch die Bearbeitungszeiten der drei Elemente zusammenaddiert werden: $\hat{y}_{(c)2} = \hat{y}_{(a)} + 2 * \hat{y}_{(b)} = 39 + 2 * 97.5 = 234$ Sekunden. Diese Eigenschaft der Formel ist von großem Vorteil, da so die Zusammenstellung von Aufgaben aus Items mit vorberechneten Itembearbeitungszeiten ohne erneute Berechnung der Einzelelemente möglich ist.

Zur Validierung der Formel wurden die Bearbeitungszeiten von 125 neuen Aufgaben an einer neuen Personenstichprobe erhoben. Der empirische Mittelwert der Bearbeitungszeit dieser Aufgaben betrug $M = 277.66$ ($SD = 91.03$), während sich die mittlere geschätzte Bearbeitungszeit auf $M = 253.79$ ($SD = 99.31$) belief, was einer Abweichung von $M_{\text{diff}} = -23.87$ ($SD_{\text{diff}} = 49.88$) beziehungsweise 8.6% entspricht.

4 Gesamtdiskussion

4.1 Zusammenfassung und Einordnung der Befunde

Gegenstand dieser Dissertation war die Auslotung und Evaluation von Optimierungsmöglichkeiten von Messinstrumenten im Large-scale Assessment, wobei die Unverfälschtheit und Validität der Messung die Optimierungskriterien darstellten. Ein großes Gefährdungspotential dieser Zielkriterien geht von Effekten aus, die im Kontext der Messung entstehen. Solche Kontexteffekte sind aus verschiedensten wissenschaftlichen Disziplinen bekannt. Im eingangs erwähnten Beispiel der Federwaage (S. 15) ist die Messung vom Schwerefeld, das an verschiedenen Orten im Universum variiert, abhängig und kann somit durchaus stark verzerrt werden. Auch bei psychometrischen Messungen treten Kontexteffekte auf. Im Gegensatz zu anderen Disziplinen sind die Natur, Größenordnung und Auftretensbedingungen von Kontexteffekten in der Psychometrie weit weniger bekannt und erforscht. Um psychometrische Messinstrumente dahingehend zu optimieren, dass diese robust gegenüber dem Auftreten von Kontexteffekten sind, müssen relevante Kontexteffekte erst einmal identifiziert werden. Deshalb widmete sich diese Dissertation der Untersuchung von drei Kontexteffekten: Testhefteffekte, Positionseffekte und Designeffekte. Im Folgendem

werden die in dieser Dissertation zu diesen Effekten gewonnenen Ergebnisse zusammengefasst und eingeordnet. Die praktischen Implikationen für die Optimierung von Testdesigns werden im nächsten Abschnitt (Abschnitt 4.2) erörtert.

Der in Daten einer Large-scale Assessment Studie vorgefundene Testhefteffekt beträgt $SD = 0.08$ Logits. Das heißt, dass das Testheft, das ein Schüler oder eine Schülerin bekommt, einen zusätzlichen Effekt (neben der Personenfähigkeit und der Itemschwierigkeit) auf die Lösungswahrscheinlichkeit ausübt. Für die meisten Testhefte lag dieser Effekt im Bereich zwischen -0.08 und 0.08 Logits. Das leichteste Testheft in dieser Studie wies eine Leichtigkeit von 0.13 Logits auf, das schwierigste -0.12 Logits. Diese Ergebnisse stehen im Einklang mit berichteten Schwierigkeiten von Testheften bestehend aus naturwissenschaftlichen Items in PISA (OECD, 2012), die zwischen -0.19 und 0.11 liegen. Die praktische Relevanz von Testhefteffekten dieser Größenordnung lässt sich an der Betrachtung der Änderung der Lösungswahrscheinlichkeiten bei Replatzierung von Items in andere Testhefte illustrieren. Wenn ein Item beispielsweise vom schwierigsten in das leichteste Testheft verschoben wird, steigt die Lösungswahrscheinlichkeit dieses Items um 5.3% an. Solche Effekte der Messinstrumente sind in der Schulleistungsforschung normalerweise unerwünscht, da viele populäre IRT-Modelle (z. B. Rasch-Modell (1PL), 2PL-, 3PL-, 4PL-Modelle) diese ignorieren. Dies kann zu Verzerrungen der Item- und/oder Personenparameter in diesen Modellen führen.

Positionseffekte wurden in den Daten des IQB-Ländervergleichs 2012 (Pant et al., 2013) bestimmt. Deren Größenordnung war ungefähr im Einklang mit jenen vorheriger Studien (z. B. Weirich, Hecht & Böhme, 2014; Robitzsch, 2009). Die standardisierten Positionseffekte (in der Parametrisierung als „Leichtigkeit“) lagen im Bereich von 0.09 Logits an der ersten Position bis -0.08 Logits an der sechsten Position. Die Varianzaufklärung durch die Positionseffekte betrug 1.8%, was Cohen's

$f^2 = 0.033$ entspricht und in den Bereich eines kleinen Effekts fällt.

Auch unterschiedliche Testdesigns können Auswirkungen auf die Akkuratheit der Messung haben. Der mittlere Bias der Itemparameter im Rasch-Modell variierte je nach verwendetem Design ($SD_{bias_\beta} = 0.029$). Zusammenfassend kann also festgehalten werden, dass die in dieser Dissertation untersuchten Kontexteffekte, die durch das Messinstrument entstehen (instrument effects nach Brennan, 1992), tatsächlich nachgewiesen werden konnten.

Die Identifikation von relevanten Kontexteffekten stellte lediglich erst die Voraussetzung für die Ableitung von Optimierungsmöglichkeiten bei der Messinstrumentekonstruktion dar. Nachdem nahegelegt war, dass Testhefteffekte tatsächlich auftreten, musste weiterhin nach dem „Warum“ gefragt werden, um neue Erkenntnisse für die Optimierung von Messinstrumenten im Large-scale Assessment zu gewinnen. Woran lag das Auftreten der gefundenen Testhefteffekte? Hierzu bot sich an, genau diejenigen Testhefteigenschaften, auf denen die Testhefte variierten, zu untersuchen. Dies ist einerseits die im Testdesign festgelegte Testheftleichtigkeit. Andererseits besteht aufgrund komplexer Randbedingungen im Testdesign üblicherweise auch Varianz in der Anzahl an Items in den Testheften. Es konnte gezeigt werden, dass die im Testdesign manipulierte Testheftleichtigkeit – operationalisiert als Mittelwert der präkalibrierten Items – tatsächlich mit der empirischen Testheftleichtigkeit zusammenhängt ($b = 0.14$). Das heißt, dass ein Testheft, das aus im Mittel ein Logit leichteren Items konstruiert wurde, 0.14 Logits leichter ist, und zwar zusätzlich zu den bereits kontrollierten Itemschwierigkeiten. Auch die Anzahl an Items hängt mit der Schwierigkeit des Testhefts zusammen. Je hinzugefügtem Item steigt die Testheftschwierigkeit um 0.014 Logits.

Dass Positionseffekte in Large-scale Assessments auftreten, konnte auch in den Daten des IQB-Ländervergleichs (Pant et al., 2013) gezeigt werden. Damit reiht sich

diese Dissertation auch in die vergleichsweise große Anzahl an Studien zum Thema Positionseffekte ein (z. B. Albano, 2013; Debeer & Janssen, 2013; Hahne, 2008; Hohensinn et al., 2008; Hohensinn, Kubinger, Reif, Schleicher & Khorramdel, 2011; Weirich, Hecht & Böhme, 2014). Wenig erforscht ist allerdings, wie gut Positionseffekte bereits im Testdesign handhabbar sind. Hierzu wird häufig die Methode der Balancierung vorgeschlagen. Diese ist besonders dann nützlich, wenn die Konstanthaltung von Kontextfaktoren nicht möglich oder nicht sinnvoll ist. Da ein Testheft normalerweise aus mehreren Items (beziehungsweise Blöcken) besteht, die nacheinander angeordnet werden, sind die Items unterschiedlich stark von Ermüdungs- und Motivationseffekten der Teilnehmerinnen und Teilnehmer betroffen. Damit jedes Item im Mittel gleich stark betroffen ist, kann das Design bezüglich Positionen balanciert werden. Das heißt, dass jedes Item (beziehungsweise jeder Block) an jeder Position gleich häufig auftritt. Wie groß der Effekt des Balancierungsgrades von Designs auf die Parameterschätzgüte von Items im Rasch-Modell ist, wurde bisher jedoch nicht quantifiziert.

Die oben beschriebene Größenordnung von Positionseffekten lässt bereits vermuten, dass der Effekt der Positionsbalance auch höchstens im kleinen Bereich zu vermuten ist. Tatsächlich konnte gezeigt werden, dass der Itembias je Punkt auf der von 0 bis 100 verlaufenden Positionsbalanceskala im Mittel um 0.001 abnimmt. Im schlechtesten Fall eines komplett unbalancierten Designs würde der mittlere Bias der Items dann 0.100 Logits betragen. Für bestimmte Items kann diese Verzerrung allerdings auch noch wesentlich höher ausfallen. Die Positionsbalancierung von Designs hat also einen Einfluss auf die Parameter in Large-scale assessments. Dieses Ergebnis ergänzt die Befunde von Frey und Bernhardt (2012), die ebenfalls Effekte bei der Verwendung von unbalancierten Designs identifizieren konnten: „Using unbalanced booklet designs can have a severe impact on population estimates of student achie-

vement in large-scale assessments.“ (S. 414). Während also bei Frey und Bernhardt Gruppenkennwerte von Personen im Fokus standen, wurde in dieser Dissertation die Auswirkung der Designbalancierung auf die Itemparameter untersucht. Da in die Bestimmung der Personenkennwerte auch die Itemparameter einfließen, ist die Akkuratheit der Itemparameter ebenfalls von großer Bedeutung.

Dass eine Positionsbalancierung bei der Verwendung des Rasch-Modells sinnvoll ist, lässt sich auch mit Hilfe der Methode *Posterior Predictive Model Checking* (PPMC; Guttman, 1967; Rubin, 1984), die von Sinharay (2005, 2006) im IRT-Kontext illustriert wurde, zeigen. Die zentrale Idee dieses bayesianischen Verfahrens liegt darin, die beobachteten Daten mit aus dem Modell vorhergesagten Daten (*replizierte Daten*) zu vergleichen. Hierzu können verschiedene im Rahmen des PPMC generierte Kennwerte herangezogen werden (siehe z. B. Sinaharay & Johnson, 2003). Sinaharay (2005) folgend gilt aber: „The preferable way to perform the posterior predictive checks is to use graphical plots (Gelman, Meng & Stern, 1996; Stern, 2000).“ (S. 376).

Wie adäquat das Rasch-Modell bei Verwendung eines balancierten Designs ist, lässt sich in Abbildung 4.1 sehen. Die Grundlage für diese Graphik bilden die Daten, die zur Bestimmung der Positionseffekte (Abschnitt 3.1.2) verwendet wurden. Die schwarze Linie kennzeichnet die Summenscore-Verteilung in diesen Daten. Zusätzlich zu diesen beobachteten Summenscores wurden basierend auf dem Rasch-Modell (obere Graphik) und dem Positionseffekte-Modell (untere Graphik) die im Rahmen der PPMC-Methode benötigten replizierten Daten generiert. Hierfür kam die Software JAGS (Plummer, 2013) in Kombination mit dem R-Paket *rjags* (Plummer, 2014) zum Einsatz. 1500 Datensätze wurden aus der posterioren Verteilung gezogen. Dadurch divergiert die vorhergesagte Summenscore-Verteilung in jedem dieser replizierten Datensätze etwas. Die graue Fläche in den Graphiken gibt jeweils das

95%-Intervall der vorhergesagten Summenscores an. Wenn das Modell eine gute Passung aufweist, sollte die beobachtete Summenscore-Verteilung (schwarze Linie) innerhalb dieses grau dargestellten Intervalls liegen. Wie man sieht, passen sowohl das Rasch-Modell als auch das Positionseffekte-Modell recht gut und quasi gleich gut. Dies ist deshalb der Fall, da ein vollständig positionsbalanciertes Testdesign verwendet wurde. Obwohl das Positionseffekte-Modell eine wesentlich geringere Deviance als das Rasch-Modell aufweist ($\text{Deviance}_{\Delta} = -7211$) und somit die Daten eigentlich besser erklärt, zeigt sich kein Unterschied bezüglich der vorhergesagten Summenscore-Verteilung, eben gerade deshalb, weil das Design balanciert war. Im Rahmen der PPMC-Methode passt das Rasch-Modell bei Verwendung eines balancierten Designs also genauso gut wie das Positionseffekte-Modell.

In unbalancierteren Designs könnte das Rasch-Modell durchaus weniger gut passen, was sich daran erkennen ließe, dass die schwarze Linie stärker außerhalb des grauen Intervalls liegt. Allerdings ist aufgrund der geringen Größe von Positionseffekten nicht zu erwarten, dass dieser Effekt graphisch allzu deutlich sichtbar wird. Da in den meisten Large-scale Assessment Studien vollständig oder nahezu vollständig positionsbalancierte Designs verwendet werden, existieren wahrscheinlich jedoch auch kaum Datensätze, an denen sich dies untersuchen ließe. Allerdings könnte hierzu auch auf empirische Daten, die unter einem balancierten Design entstanden sind, zurückgegriffen werden. Durch gezieltes Löschen von Beobachtungen (z. B. Löschen bestimmter Testhefte) könnte die Balancierung des Designs reduziert werden. Diese Methode wurde beispielsweise in Frey und Bernhardt (2012) verwendet, um den Effekt zwischen einem balancierten Youden Square Design und einem unbalancierten Design auf zentrale Ergebnisse in PISA zu untersuchen. Bezüglich der Auswirkung der Designbalancierung auf den Modellfit wird festgestellt: „Unbalanced booklet designs cannot necessarily be detected by a lack of model fit ... “ (S. 414). Diese Aussage

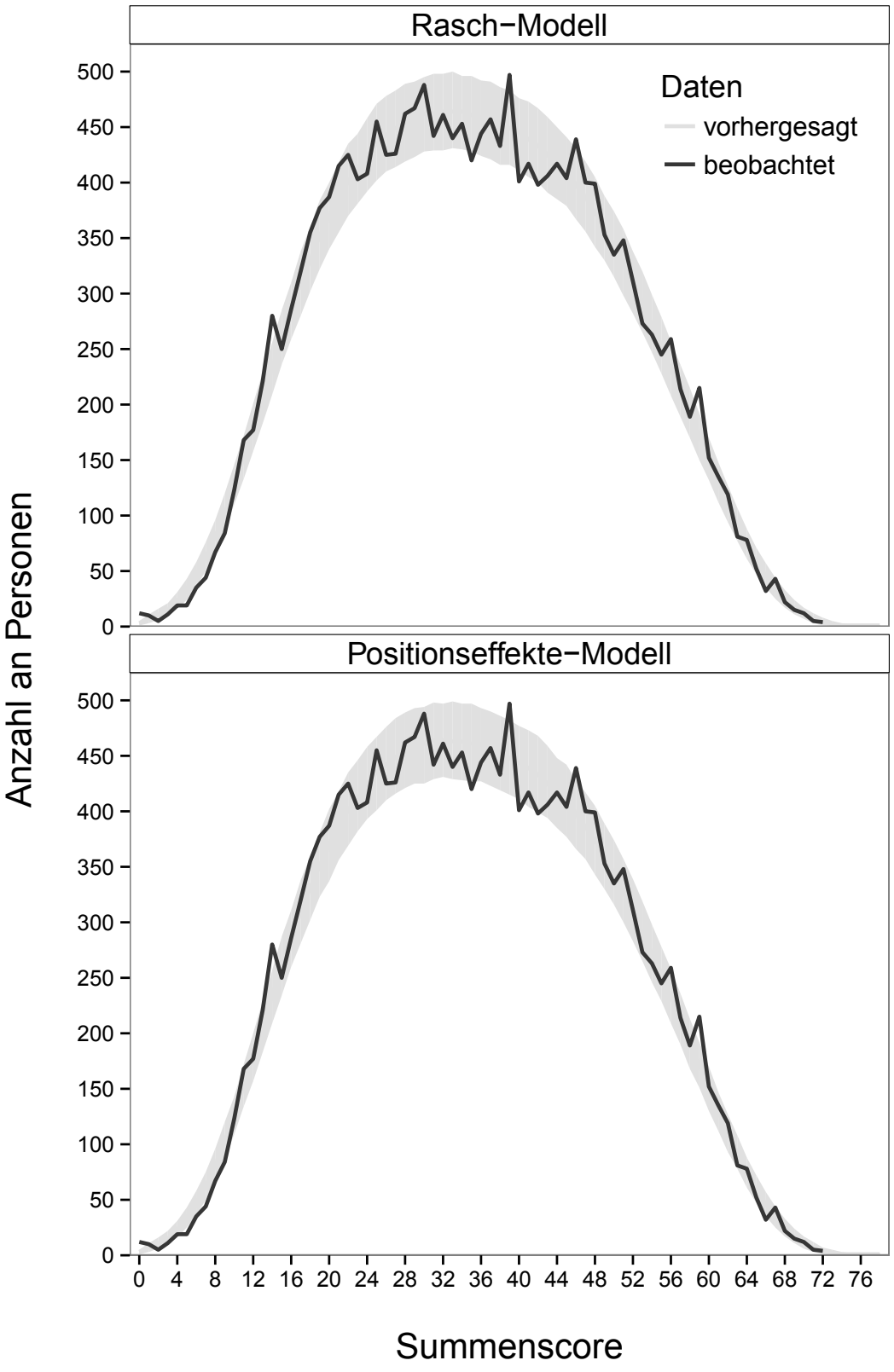


Abbildung 4.1: Beobachtete und vorhergesagte Summenscore-Verteilung im Rahmen des posterior predictive model checking

bezieht sich allerdings auf die Verwendung der Deviance als Modellpassungskriterium.

Warum könnte ein bezüglich Positionen unbalancierteres Design bei Verwendung des Rasch-Modells sogar einen besseren Modellfit aufweisen? Frey und Bernhardt (2012) führen hierzu aus: „... a better model fit can sometimes be expected for unbalanced designs than for balanced designs if an unbalanced design systematically diminishes effects violating the assumptions of the IRT model used for scaling.“ (S. 412). Für den konkreten Fall der Verwendung des Rasch-Modells in Daten mit Positionseffekten könnte dies folgendermaßen verstanden werden: In positionsbalancierten Designs existiert theoretisch für jedes Item an jeder Position ein Parameter, wobei sich die Parameter eines Items jeweils aufgrund der Positionseffekte unterscheiden. Anders formuliert, es gibt itemspezifische Varianz. Diese wird im Rasch-Modell nicht modelliert. In unbalancierten Designs hingegen ist diese Varianz kleiner oder null. Wenn jedes Item beispielsweise nur an einer Position vorkommt, gibt es auch nur einen Itemparameter. Dadurch sind die Itemparameter zwar vollständig mit den Positionen konfundiert, da aber keine positionsabhängige Itemvarianz existiert und damit durch das Rasch-Modell ignoriert werden kann, könnte der Modellfit unter Verwendung eines unbalancierten Designs besser als unter Verwendung eines balancierten Designs sein.

Insgesamt sind also verschiedene Szenarien zum Zusammenhang zwischen dem Modellfit und der Designbalancierung denkbar, die auch von der verwendeten Methode zur Beschreibung des Modellfits abhängen.

Eine weitere Balancierung, die häufig in Testdesigns von Large-scale Assessment Studien vorgenommen wird, ist die Blockpaarbalancierung. Für diese Balancierung zeigte sich ein Nulleffekt. Folglich ist die relative Anzahl an realisierten Blockpaaren irrelevant für die Itemparameterschätzung. Diese Aussage gilt jedoch nur für die

Annahmen, die in dieser Dissertation über den Blockpaareffekt getroffen wurden (siehe Abschnitt 4.3).

Neben Kontexteffekten, die durch die spezifische Konstruktion des Messinstruments entstehen, widmete sich diese Dissertation auch noch einem weiteren, zur Konstruktion von Testheften relevanten, Kriterium, den im Testdesign festgelegten Testheftbearbeitungszeiten. Die akkurate Setzung und Einhaltung einer definierten Sollbearbeitungszeit des Testhefts ist für die Durchführung und Validität von Large-scale Assessments von großer Wichtigkeit. Einerseits bleibt bei Überschätzung der Bearbeitungzeit wertvolle Testzeit ungenutzt, da die Testteilnehmerinnen und -teilnehmer früher als eingeplant den Test beenden. Andererseits treten bei Unterschätzung der Bearbeitungszeit *speededness* Effekte auf, die die Validität von Power-Tests negativ beeinflussen können (Lu & Sireci, 2007). Zur akkuraten Bestimmung der Testheftbearbeitungszeit ist die Kenntnis der Item- beziehungsweise Aufgabenbearbeitungszeiten notwendige Voraussetzung. Diese können beispielsweise in Vorstudien bestimmt oder von Experten geschätzt werden. Als kostengünstige, einfache Möglichkeit bietet sich aber auch deren Vorhersage aus leicht verfügbaren Aufgabeneigenschaften an. Anhand einer empirischen Stichprobe konnte eine Formel abgeleitet werden (Gleichung 3.1 auf S. 68), mit deren Hilfe Bearbeitungszeiten aus Aufgabeneigenschaften berechnet werden können.

Die Ergebnisse sind plausibel und vergleichbar mit jenen aus den wenigen anderen Studien zu diesem Thema. Die Anzahl an Wörtern war ebenfalls ein relevanter Prädiktor in den Studien von Halkitis et al. (1996), Bergstrom et al. (1994) und Swanson et al. (2001). Der gefundene Effekt der Wortanzahl war mit 0.39 s ähnlich zu dem in Swansons Studie (0.5 s). Das als relevanter Prädiktor identifizierte Antwortformat wurde in vorherigen Studien nicht variiert, womit auch keine Befunde zum Vergleich vorliegen. Dass Multiple-Choice-Items schneller als Items mit offenem

Antwortformat bearbeitet werden können, ist jedoch ein sehr plausibles Ergebnis. Während Aufgabeneigenschaften einen hohen Anteil (94.3%) an der Variation der Aufgabenbearbeitungzeiten erklären konnten, zeigten sich für die untersuchten Personeneigenschaften Nulleffekte. Dieser Befund ist konsistent mit jenen aus Bergstrom et al. (1994): „examinee characteristics are generally not related to response time“ (S. 13). Weiterhin erlaubte das verwendete statistische Verfahren (Linear Mixed Models) die Modellierung von Aufgaben × Personen-Interaktionen. Hierbei zeigte sich, dass Mädchen und Gymnasiasten signifikant länger an Items mit offenem Antwortformat arbeiteten.

Die in dieser Dissertation gewonnenen Ergebnisse zu Testhefteffekten, Positionseffekten, Designeffekten und Aufgabenbearbeitungszeiten können verwendet werden, um Messinstrumente zu optimieren. Diese Optimierungsmöglichkeiten werden im nächsten Abschnitt erörtert.

4.2 Praktische Implikationen

Ein Hauptanliegen dieser Dissertation war, Erkenntnisse zur Optimierung von Messinstrumenten zu erlangen. Deshalb können auch aus jedem der drei Beiträge konkrete praktische Implikationen abgeleitet werden. Diese sind primär für Praktikerinnen, Praktiker, Wissenschaftlerinnen und Wissenschaftler im Bildungsforschungsbereich relevant, die Messinstrumente im Large-scale Assessment erstellen und einsetzen wollen.

Das erste zentrale Ergebnis ist: Testheft-, Positions- und Designeffekte können in Large-scale assessement Studien auftreten. Die in den untersuchten Datensätzen gefundenen Effekte sind zwar klein, sollten aber dennoch nicht ignoriert werden.

Was kann zu deren Handhabung also getan werden? Diese Frage wird mit Hinblick auf die in Abschnitt 1.3.4 dargestellten Methoden im Folgenden erörtert.

Zur Handhabung von Testhefteffekten können Testhefte im statistischen Auswertungsmodell berücksichtigt werden, was dazu führt, dass die Itemparameter im Modell nicht durch Ignorieren der Testhefteffekte verzerrt werden. Diese Strategie verfolgt beispielsweise auch PISA (OECD, 2009): „When estimating the item parameters, booklet effects were included in the measurement model to prevent confounding item difficulties and booklet effects. For the ConQuest model statement, the calibration model was: item + item*step + booklet." (S. 155). Dieses Modell entspricht konzeptuell dem in dieser Dissertation im GLMM Framework postulierten Testheftemodell in Gleichung 2.1 (bis auf den Term item*step, der auf die Verwendung polytomer statt dichotomer Daten – und damit auch auf ein anderes Messmodell – rekurriert). Durch welche Eigenschaften der Testhefte die Testhefteffekte verursacht wurden, war jedoch nicht Gegenstand der Untersuchung in PISA. Hierüber wurde lediglich spekuliert: „... due to the different location of domains within each of the booklets it was expected that there would ... be booklet influences on the estimated proficiency distributions." (S. 155). Zur Handhabung von Testhefteffekten durch Modellierung ist die Kenntnis deren Ursachen aber auch nicht nötig. Es ist ausreichend, die Testhefte (und ggf. Designvariablen, von denen die Testheftzuteilung abhängig war) im Modell zu integrieren.

Durch diese Integration wird jedoch erst einmal nur sichergestellt, dass die Itemparameter nicht durch das Ignorieren von Testheftparametern verzerrt werden. Im nächsten Schritt sollten auch noch die Personenparameter korrigiert werden (OECD, 2009): „The booklet effects are the amount that must be added to the proficiencies of students who respond to each booklet." (S. 221). Dies kann durch manuelles Addieren der im *calibration model* bestimmten Testhefteffekte zu den im *population*

model bestimmten Personenparametern erfolgen. Alternativ wäre auch denkbar, die vorab bestimmten Testheftparameter in das *latent regression model* (Wu, Adams, Wilson & Haldane, 2007), das in der deutschsprachigen Literatur oft auch *Hintergrundmodell* (z. B. Weirich, Haag & Roppelt, 2012) genannt wird, zu integrieren. Damit würden die in diesem Modell bestimmten Personenkennwerte (insbesondere auch die häufig verwendeten *plausible values*) für die Testhefteffekte korrigiert werden. Mit „korrigiert“ ist hier gemeint, dass die Personenkennwerte dann bereits die Testhefteffekte beinhalten, womit eine manuelle Korrektur nicht mehr nötig wäre.

Die Gründe, warum Personenparameter korrigiert werden sollten, werden jedoch selten expliziert. Im technischen Bericht von PISA 2000 (OECD, 2002) wird lediglich ausgeführt, dass das Addieren der Testhefteffekte zu den Personenparametern die Mittelwerte der Länder – und damit auch das Länder-Ranking – nicht verändert. Hier zeigt sich auch wieder der Unterschied zwischen den verschiedenen Ebenen im Large-scale Assessment. Während Testhefteffekte auf Individualebene wirken, werden bestimmte Kennwerte auf Gruppenebene nicht beeinflusst. Deshalb sind Testhefteffekte in bestimmten Anwendungsfällen unproblematisch. Welche Gruppenkennwerte unter welchen Bedingungen und Testdesigns von Testhefteffekten beeinflusst werden, ist bisher allerdings nicht hinreichend geklärt. Frey und Bernhardt (2012) beschreiben ein Szenario, unter dem die Testhefteffektekorrektur auch für Gruppenkennwerte hilfreich sein könnte: „... applying the booklet correction ... avoids potential problems when analyzing small subsamples where a uniform distribution of booklets might not be guaranteed due to small sample sizes.“ (S. 399).

Ebenfalls sollten die in Abschnitt 1.3.1 bereits andiskutierten politischen Implikationen der Verwendung verschiedener Testhefte beachtet werden. Die Testhefteffektekorrektur stellt konzeptuell eine Bonus-Malus-Modifikation dar. Testteilnehmer, die überdurchschnittlich schwere Testhefte bearbeitet haben, werden „belohnt“

(Bonus), während Testteilnehmer mit leichten Testheften „bestraft“ werden (Malus). Für Entscheidungen auf Individualebene könnte es deshalb durchaus herausfordernd sein, solche Korrekturmethoden gegenüber bestimmten Interessen- und Anspruchsgruppen zu erklären und zu rechtfertigen, zumal auf komplexere statistische Verfahren oft mit Skepsis und Reserviertheit reagiert wird. Deswegen sollten für Vergleiche von Personen keine unterschiedlichen Testhefte verwendet werden. Aber auch für die im Large-scale Assessment angestrebten Gruppenvergleiche gilt teilweise diese Argumentation. Wenn durch die Anpassung von Testheften an bestimmte Personengruppen eine Testhefteffektekorrektur nötig wird, könnte auch hier Erklärungsbedarf zur verwendeten Methodik entstehen. Weiterhin tangiert die Testheftanpassung das weite Feld der Testfairness, wobei Brennan (1992) zwei mögliche Positionen zu diesem Thema zusammenfasst: „ ... under the traditional definition of standardized testing, all examinees are exposed to the same measurement experience. If everyone has been treated equally, so the argument goes, then everyone has been treated fairly. On the other hand, many have pointed out the fallacy of considering equal treatment to be fair treatment.“ (S. 235). Da das Thema Testfairness nicht Gegenstand dieser Dissertation ist und weitere Ausführungen den Rahmen sprengen würden, sei hier lediglich auf weiterführende Literatur verwiesen (z. B. Cole & Zieky, 2001; Joint Committee on Testing Practices, 2004; Camilli, 2006; Nachtigall, Kröhne, Enders & Steyer, 2008).

Sollen Testhefteffekte bereits im Testdesign berücksichtigt werden, können deren Eigenschaften, die für die Effekte verantwortlich sind, konstant gehalten werden. Bezüglich der beiden untersuchten Testhefteigenschaften hieße das also, die Schwierigkeit der Testhefte und die Anzahl an Items in den Testheften konstant zu halten. Dies schränkt allerdings die Generalisierbarkeit der Ergebnisse ein. Ergebnisse, die unter Verwendung von Testheften mit verschiedenen Ausprägungen auf den Ei-

genschaften entstanden sind, sind nicht vergleichbar. In Testdesigns für zukünftige Studien sollte deshalb dann auch weiterhin dieselbe Anzahl an Items und dieselbe Testheftschwierigkeit verwendet werden.

Durch konstante Testheftschwierigkeiten gibt man jedoch die Möglichkeit auf, fähigkeitsdivergente Teilstichproben mit schwierigkeitsangepassten Testheften zu versorgen, was aus Motivationsgründen oder statistischen Gründen (z. B. Messpräzision) sinnvoll sein könnte. Bei diesem Vorgehen ist jedoch zu beachten, dass Testhefte nicht verlinkt werden können. Dies hat zur Folge, dass Testhefte, die ausschließlich bestimmten Personengruppen zugewiesen werden, in Referenz zu deren Fähigkeit und nicht bezogen auf die Gesamtstichprobe geschätzt werden. Testhefte, die kompetenteren Schülerinnen und Schülern zugewiesen werden, erscheinen dann leichter und Testhefte, die weniger kompetente Schülerinnen und Schüler bekommen, schwerer. Um diesem Problem zu begegnen, sollte die Designvariable, die die nichtzufällige Zuweisung der Testhefte beschreibt, im Modell berücksichtigt werden (siehe Abschnitt 2.1.1). Sollen beispielsweise Gymnasiasten im Mittel schwierigere Testhefte bekommen als Schüler und Schülerinnen anderer Schularten, sollte diese Gruppenvariable auch Eingang in das Messmodell finden. Diese Empfehlung ist im Übrigen nicht nur für die Bestimmung der Testhefteffekte, sondern auch für die Schätzung der Itemparameter bei Verwendung der *Marginal Maximum Likelihood* (MML) Methode relevant (DeMars, 2002).

Die Handhabung von Testhefteffekten im Testdesign setzt voraus, dass die Testhefteigenschaften, die die Testhefteffekte erzeugen, bekannt sind. Die beiden in dieser Dissertation als relevant identifizierten Testhefteigenschaften sind wahrscheinlich nur zwei von potentiell vielen. Das heißt, dass auch trotz Berücksichtigung von bestimmten Testhefteigenschaften im Testdesign (z. B. durch Konstanthaltung) Testhefteffekte (die auf anderen, nicht berücksichtigten Testhefteigenschaften beruhen)

auftreten können. Es ist deshalb ratsam, Testhefteffekte in jeder Anwendung zu untersuchen. Somit kann eruiert werden, wie groß die Testhefteffekte tatsächlich sind und ob diese ignoriert werden können. Zur Frage, ab welcher Größenordnung von Testhefteffekten deren Ignorieren zu Problemen führt, gibt es jedoch keine gesicherteten Erkenntnisse. Deshalb ist zum jetzigen Stand der Forschung die sicherste Strategie, Testhefte immer mitzumodellieren.

Die Methode der Konstanthaltung eignet sich auch zur Handhabung von Positionseffekten nur bedingt. Natürlich könnte jedes Item in allen Studien immer an derselben Position präsentiert werden. Dadurch wäre die Itemschwierigkeit zwar trotzdem durch die Schwierigkeit der Position konfundiert. Da diese Konfundierung aber in allen Studien gleich (oder zumindest ähnlich) wäre, gäbe es bezüglich der Vergleichbarkeit der Ergebnisse eher keine Probleme. Die Nachteile dieses Vorgehens liegen aber darin, dass (a) die Kenntnis der „wahren“ Itemschwierigkeit aus inhaltlicher Sicht relevant sein könnte und dass (b) wichtige Flexibilität bei der Erstellung von Testdesigns aufgegeben würde. Beispielsweise ist die *Bookmark Method* (Mitzel et al., 2001) zur Konstruktion von Kompetenzstufen darauf angewiesen, dass die Items nach ihrer tatsächlichen Schwierigkeit angeordnet werden können. Anders formuliert, man benötigt hier nicht die Itemschwierigkeit an einer bestimmten Position im Test, sondern die generelle positionsunabhängige Itemschwierigkeit, um inhaltlich valide Kompetenzstufen ableiten zu können. Technisch gesehen bedeutet eine konstante Itemposition eine sehr hohe Restriktion, die im Testdesign quasi nicht umsetzbar ist. Items könnten dann nur noch Testheften zugeordnet werden, wenn die jeweilige erforderliche Position vakant ist. Dies schränkt die Auswahl an Items aus dem Itempool stark ein, da immer nur so viele Items, die eine bestimmte Position erfordern, aus dem Pool verwendet werden können, wie diese bestimmte Position im Testdesign verfügbar ist. Aus diesen Gründen bietet sich die Konstanthaltung der

Itempositionen nicht an.

Analog zu Testhefteffekten, könnten Positionseffekte auch durch Berücksichtigung im statistischen Modell kontrolliert werden. Allerdings gelten hier auch die Nachteile der Modellierungsstrategie (siehe Abschnitt 1.3.4). Stattdessen hat sich im Large-scale Assessment Kontext die Methode des Balancierens der Items bezüglich der Positionen etabliert. Hierbei werden Items (bzw. Blöcke) so auf Positionen verteilt, dass diese gleich häufig auf allen Positionen auftreten. Dadurch mitteln sich Positionseffekte heraus. Wie effektiv ist diese Strategie allerdings? Es konnte gezeigt werden, dass der Itembias im Mittel um 0.001 Logits sinkt, je balancierter das Design (auf einer Skala von 0 bis 100) ist, wobei das balancierteste Design mit einer Balance von 100 quasi biasfrei ist. Es ist also angeraten, Positionen so gut wie möglich zu balancieren. Allerdings ist durch die geringe Größe von Positionseffekten auch der Effekt der Balancierung klein. Man könnte auch durchaus überlegen, geringfügig weniger balancierte Designs zu verwenden, wenn andere konfligierende Designeigenschaften prioritärer zu bewerten sind.

In Testheften (mit mehr als einem Block) werden immer auch bestimmte Paare an Blöcken miteinander kombiniert. Auch dieses Auftreten an Blockpaaren kann balanciert werden, indem alle Blockpaare gleich häufig im Design vorkommen. Diese Balancierung hatte keine Auswirkungen auf die Schätzgüte der Itemparameter (unter den Annahmen, die über die Art des Blockpaareffektes getroffen wurden, siehe Abschnitt 4.3). Dies ist für Testdesigner ein durchaus positiv zu bewertender Befund, da die Blockpaarbalance relativ beliebig reduziert werden kann, ohne negative Effekte erwarten zu müssen. Dies ermöglicht auch, Designs zu verwenden, die komplett positionsbalanciert, aber nur nahezu blockpaarbalanciert sind. Solche Designs entstehen häufig bei dem Versuch, Youden Square Designs zu konstruieren. Die Ergebnisse legen nahe, dass deren Verwendung unproblematisch ist. Dies ist eine sehr

hilfreiche Erkenntnis, da viele große Youden Square Designs nicht existieren oder noch nicht gefunden wurden.

Auch für das Vorhersagemodell für Aufgabenbearbeitungszeiten liegt die praktische Relevanz klar auf der Hand: Die Formel 3.1 auf Seite 68 kann direkt verwendet werden, um für Aufgaben, Aufgabenstimuli und Items Bearbeitungszeiten aus den leicht verfügbaren Aufgabenmerkmalen Itemanzahl, Anzahl Wörter und Antwortformat (Multiple-Choice, Kurzantwort, Erweiterte Antwort) zu berechnen (siehe auch die Beispielrechnungen in Abschnitt 3.2.3). Damit kann auf kostspielige Pilotstudien oder potentiell inakkurate Expertenschätzungen zur Bestimmung der Bearbeitungszeit verzichtet werden. Der gefundene Nulleffekt von Personeneigenschaften und der Bearbeitungszeit ist ebenfalls eine gute Nachricht für Testdesigner, da so für unterschiedliche Gruppen (z. B. Schülerinnen und Schüler verschiedener Schularten) keine unterschiedlichen Testformen erstellt werden müssen. Die beiden signifikanten Aufgaben × Personen-Interaktionen könnten aber herangezogen werden, um bei Verwendung bestimmter Aufgaben in bestimmten Gruppen die Bearbeitungszeit anzupassen. Sollen zum Beispiel in Gymnasien Items mit dem Antwortformat „Erweiterte Antwort“ eingesetzt werden, müsste mit einem höheren Zeitbedarf kalkuliert werden. Bei Anpassung der Bearbeitungszeit an bestimmte Personengruppen sollten allerdings immer auch sorgfältig die politischen Konsequenzen mit in Betracht gezogen werden. Gruppen, die miteinander verglichen werden sollen, sollte aus Gründen der Fairness und Vergleichbarkeit auch die gleiche Bearbeitungszeit zur Verfügung gestellt werden.

4.3 Methodische Bewertung und Grenzen der Arbeit

Während im vorherigen Abschnitt die praktischen Implikationen dargestellt wurden, sollen die Befunde jetzt kritisch betrachtet und bewertet werden. Es wird erörtert, wie gerechtfertigt und belastbar die Ergebnisse und die aus diesen abgeleiteten Handlungsempfehlungen sind und welche Aussagen nicht getroffen werden können.

Die im Rahmen der Untersuchung von Testhefteffekten analysierten Testhefteigenschaften waren die im Design festgelegte Testheftschwierigkeit und die Anzahl an Items im Testheft. Um deren Einfluss auf die tatsächliche empirische Testheftschwierigkeit zu untersuchen, wurden diese beiden Testhefteigenschaften variiert. Es lag also im Prinzip ein experimentelles Design vor: Stimulusmaterial wurde variiert und dann zufällig beziehungsweise bedingt zufällig an die Testteilnehmerinnen und -teilnehmer verteilt, wobei die Bedingung die Schulart war. Gymnasiasten bekamen im Mittel etwas schwerere Testhefte als Schülerinnen und Schüler anderer Schularten. Innerhalb der Schulart (Gymnasium vs. andere Schularten) war die Zuweisung der Testhefte zufällig.

Können die Zusammenhänge zwischen den beiden Testhefteigenschaften und der Testheftschwierigkeit als kausal interpretiert werden? Um eine kausale Interpretierbarkeit zu gewährleisten, müssen drei Voraussetzungen gelten (Suppes, 1970):

1. Zwischen der Ursache und dem Outcome muss ein Zusammenhang bestehen.
2. Die Ursache geht der Wirkung voraus.
3. Der Zusammenhang zwischen Ursache und Wirkung bleibt auch nach Kontrolle von Drittvariablen bestehen.

Die ersten beiden Punkte sind erfüllt. Zwischen den Testhefteigenschaften und der Testheftschwierigkeit besteht ein Zusammenhang, wie im NEGBP Modell gezeigt wurde (Abschnitt 3.2.1). Weiterhin wurden die Testhefteigenschaften vor der Erhebung manipuliert. Kann jedoch das Wirken von Drittvariablen ausgeschlossen werden? Der Goldstandard in experimentellen Designs zur Kontrolle von Drittvariablen ist die Randomisierung. Die Zuweisung der Testhefte zu Personen erfolgte bedingt zufällig, wobei die Bedingung (Schulart) im Modell berücksichtigt wurde. Es kann also davon ausgegangen werden, dass die Ergebnisse eher nicht durch Personenvariablen konfundiert sind. Die Zuweisung der Items zu Testheften erfolgte allerdings nicht zufällig, sondern in Abhängigkeit ihrer Schwierigkeit, gerade genau um Testhefte mit unterschiedlicher Schwierigkeit zu erzeugen. Damit war die Wahrscheinlichkeit, dass schwierigere Items in leichteren Testheften und leichtere Items in schwierigeren Testheften platziert werden, geringer als im Durchschnitt. Dadurch könnten Item- und Testhefteffekte konfundiert sein. Diese Konfundierung ist allerdings nicht auflösbar, wenn man die Testheftschwierigkeit auf Basis der Itemschwierigkeiten konzipiert. Die Konstruktion von Testheften durch eine zufällige Auswahl an Items hätte nämlich zur Folge, dass die Testhefte quasi gleich schwer wären. Auch die Anzahl an Items im Testheft war ebenfalls nicht zufällig und ist potentiell mit Drittvariablen konfundiert. Zum Beispiel könnte die Itemlänge vom Kompetenzbereich (Biologie, Chemie, Physik) abhängen. Dadurch besitzen Testhefte mit überproportionalem Anteil eines bestimmten Kompetenzbereiches dann auch über- oder unterdurchschnittlich viele Items, womit die Itemanzahl im Testheft nicht mehr unabhängig vom Kompetenzbereich wäre. Das Wirken von konfundierenden, instrumentenseitigen Drittvariablen kann also nicht ausgeschlossen werden.

Warum wurde die Testheftschwierigkeit und die Itemanzahl untersucht? Die Auswahl genau dieser Testhefteigenschaften beruht auf deren Relevanz beziehungsweise

Potenzial für das Testdesign. Bezüglich der Anzahl an Items wird normalerweise nicht versucht, vorsätzlich Varianz zu erzeugen; vielmehr ergibt sich diese Varianz im Testdesign aufgrund anderer Restriktionen, deren Erfüllung höhere Priorität als das Konstanthalten der Itemanzahl hat. Dies stellt ein häufig auftretendes Phänomen in Testdesigns im Large-scale Assessment dar. Deshalb war es wichtig, zu untersuchen, wie groß der Einfluss der Itemanzahl auf die Testheftschwierigkeit ist. Wäre dieser null, müsste auf die Itemanzahl keine Rücksicht genommen werden. Es gäbe dann also mehr Spielraum bei der Erstellung der Testhefte. Die Ergebnisse legen allerdings nahe, dass mit steigender Itemanzahl auch die Testheftschwierigkeit steigt. Zur Vermeidung dieser Effkete könnten diese Testhefteigenschaften im Testdesign konstant gehalten werden.

Die Variation der Testheftschwierigkeiten könnte jedoch auch erwünscht sein, um für bestimmte Personengruppen mit unterschiedlicher Fähigkeit schwierigkeitsangepasste Testhefte zu erzeugen. Hierdurch könnte die Teilnahmemotivation steigen, da die Teilnehmer nicht unter- oder überfordert, sondern optimal beansprucht werden. Dies ist auch die zentrale Idee des adaptiven Testens (z. B. Linacre, 2000; van der Linden & Glas, 2010; Wainer, 2000). Ein weiterer Vorteil von schwierigkeitsangepassten Testheften ist die höhere mittlere Messpräzision, da in IRT-Modellen die Messpräzision am höchsten ist, wenn die Itemschwierigkeit mit der Personenfähigkeit identisch ist. Die Untersuchung dieser potentiellen Vorteile war allerdings nicht Gegenstand dieser Dissertation. Deshalb können dazu keine Aussagen gemacht werden. Weiterhin können auch keine Aussagen darüber getroffen werden, wie stark die Parameterverzerrung in Modellen ist, die Testhefteffekte nicht mitmodellieren (z. B. im Rasch-Modell). Deshalb sollte eigentlich in jeder Studie untersucht werden, ob Testhefteffekte auftreten und gegebenenfalls berücksichtigt werden müssen.

Im Gegensatz zu Testheft- und Positionseffekten, die in empirischen Daten un-

tersucht wurden, kam zur Analyse von Designeffekten eine Simulation zum Einsatz, da empirische Datensätze, deren Designs auf den im Fokus stehenden Designeigenschaften variieren, nicht existieren beziehungsweise die Erhebung solcher Daten nur schwer vertretbar ist. Wenn man nämlich annimmt, dass bestimmte Ausprägungen auf bestimmten Designeigenschaften die Akkuratheit der Messung negativ beeinflussen, würde man solche Designs in der Praxis nicht einsetzen. Deshalb bleibt für Fragestellungen nach dem Einfluss von Designeigenschaften nur der Rückgriff auf simulierte Daten (oder auf Designmodifikationen in empirischen Datensätzen).

In der durchgeführten Simulation wurde die Güte der Parameterschätzung in verschiedenen Bedingungen untersucht. Dabei sind bestimmte Aspekte bezüglich des Simulationsmodells und der Gestaltung der Simulationsbedingungen zu beachten. Der Wahl des Simulationsmodells kommt eine entscheidende Bedeutung zu, da nur die Effekte untersucht werden können, die auch in den durch das Simulationsmodell generierten Daten vorhanden sind. Da Positionseffekte ein gut dokumentiertes Phänomen darstellen, ist auch das Simulationsmodell schlüssig: Zusätzlich zu den Item- und Personenparametern hängt die Lösungswahrscheinlichkeit ebenfalls von der Position im Testheft ab. Die Größe dieser Positionseffekte kann aus der Analyse empirischer Daten gewonnen werden. Dazu wurden Daten aus dem IQB-Ländervergleich 2012 (Pant et al., 2013) herangezogen, wobei die Ergebnisse mit denen aus anderen Studien gut übereinstimmten. Auf Grundlage dieser empirisch ermittelten Positionseffekte wurden dann simulierte Daten generiert. In diesen ist also der Positionseffekt enthalten. Weil untersucht werden sollte, wie stark die Parameterverschätzung im Rasch-Modell in Abhängigkeit der Balancierung des Designs ist, wurde genau diese Balance der Positionen variiert. Hierzu bedarf es eines Kennwertes, der die Positionsbalance beschreibt. Aufgrund der eher spärlichen Literaturlage wurde ein neuer Kennwert auf Basis der Varianz-Kovarianz-Matrix des

Designs vorgeschlagen (Abschnitt 2.2.2), der im Bereich von 0 (unbalanciert) bis 100 (balanciert) rangiert. Obwohl die Konstruktion dieses Kennwerts eine große Plausibilität aufweist, sollte dieser mit Vorsicht verwendet werden, da dessen Eigenschaften bisher unzureichend untersucht sind. Mithilfe dieses Kennwerts war es möglich, die Positionsbalance verschiedener Designs zu beschreiben und Designs mit spezifischer Positionsbalance zu konstruieren.

Während das Auftreten von Positionseffekten ein bekanntes Phänomen darstellt, verhält es sich für Effekte der Blockpaare anders. Hier ist unklar, wie sich die Realisation von bestimmten Blockpaaren im Design auswirkt. Auch werden in der Literatur keine Gründe zur Notwendigkeit der Blockpaarbalancierung angegeben. Trotzdem werden Blockpaare in Large-scale Assessment Programmen (z. B. in PISA, OECD, 2012; in NAEP, Allen, Donoghue & Schoeps, 2001; im IQB-Ländervergleich, Hecht, Roppelt & Siegle, 2013) routinemäßig balanciert. Auf Grundlage der fehlenden Evidenz zur Natur von Blockpaareffekten wurde angenommen, dass die Balancierung von Blockpaaren für die Stabilität der Schätzalgorithmen wichtig ist. Mit sinkender Balancierung vermindert sich nämlich auch die zur Schätzung zur Verfügung stehende Itemkovarianzinformation. In anderen Worten, der Anteil an Missings in der Itemkovarianzmatrix steigt. Im Gegensatz zu Positionseffekten wird damit der Blockpaareffekt nicht als psychologisch-methodisches, sondern als statistisches Problem konzipiert. Deshalb braucht auch kein Blockpaareffekt in das Simulationsmodell integriert werden. Die Interpretation der Ergebnisse muss jedoch im Einklang mit der gewählten Konzeptualisierung stehen. Der Nulleffekt der Blockpaarbalance kann nur dahingehend interpretiert werden, dass das eingesetzte statistische Verfahren (Rasch-Modell mit einem MML-Schätzer) robust gegenüber geringer Itemkovarianzinformation ist. Die Kovarianzinformation kann also fast beliebig reduziert werden, ohne dass sich die Güte der Itemparameterschätzung verringert. Eine an-

dere Interpretation ist im verwendeten konzeptuellen Rahmen nicht möglich. Ein alternatives Konzept wäre – ähnlich wie bei den Positionseffekten – einen psychologischen Effekt anzunehmen. Beispielsweise könnte die Bearbeitung von Items nach bestimmten anderen Items zusätzlich noch schwerer sein, weil nach einem besonders schweren Item die Motivation, das nächste Item zu lösen, stark sinkt. Solche Effekte werden unter dem Begriff *carryover effects* zusammengefasst (z. B. Frey, Hartig & Rupp, 2009). Können Blockpaareffekte als Carryover-Effekte verstanden werden? Im Prinzip nein. Bei der in Large-scale Assessment Studien üblichen Balancierung von Blockpaaren kommt es lediglich darauf an, alle Blockpaare mit gleicher Häufigkeit im Testdesign zu realisieren. Die Reihenfolge der beiden Blöcke, die ein Blockpaar bilden, wird dabei nicht beachtet. Es kann also Block 1 auf Block 2 folgen, oder andersherum – in beiden Fällen ist das Blockpaar realisiert. Außerdem müssen die Blöcke nicht direkt aufeinanderfolgen; es können auch ein oder mehrere Blöcke dazwischen liegen. Aus diesen Gründen ist die Annahme der Blockpaareffekte als Carryover-Effekte unplausibel und die Blockpaarbalancierung keine Balancierung bezüglich Carryover-Effekten.

Ein eher selten diskutiertes Problem, das in logistischen Regressionen beziehungsweise Modellen mit nichtlinearer Link-Funktion – und damit auch in vielen IRT-Modellen – auftritt, betrifft die modellabhängige Veränderung der Effektgrößen durch relevante, aber nicht modellierte Faktoren (siehe auch Abschnitt 1.3.6). Mood (2010) stellt fest: „... the problem of unobserved heterogeneity has escaped the attention of the large majority of users of logistic regression“ (S. 73). Da die Fehlervarianz in logistischen Modellen fixiert ist, steigt die Varianz der latenten Skala bei Hinzufügen von Faktoren, die die Varianzaufklärung erhöhen. Um diesem Problem zu begegnen, könnten die Effekte an $SD = 1$ der latenten Skala standardisiert werden (z. B. Winship & Mare, 1984; Mood, 2010). Dieses Vorgehen wur-

de zur Standardisierung der Positionseffekte verwendet. Damit werden die Effekte aus Modellen mit verschiedenen Faktoren in der gleichen Stichprobe vergleichbar. Zum Vergleich zwischen verschiedenen Personengruppen oder Stichproben ist diese Standardisierung aber nur bedingt aussagekräftig, nämlich nur unter der Annahme, dass der Effekt der nicht modellierten Faktoren für die Gruppen beziehungsweise Stichproben äquivalent ist. Außerdem bleibt die „wahre“ Skala auch weiterhin unbekannt, da nicht auszuschließen ist, dass zusätzliche relevante Faktoren existieren, die nicht im Modell berücksichtigt wurden. Trotzdem ist die Standardisierung von Effekten in IRT-Modellen ein Schritt in die richtige Richtung, um dem Problem der modellabhängigen Effektnormierung zu begegnen und zur Vergleichbarkeit von Effekten beizutragen. Neben der Standardisierung werden von Mood (2010) weitere Vorgehensweisen vorgeschlagen, die in zukünftigen Studien zu Positionseffekten Anwendung finden könnten. Insgesamt sollte dem Problem der modellabhängigen Effektnormierung im Large-scale Assessment mehr Aufmerksamkeit als bisher zuteil werden.

Auch bei der Erstellung des Vorhersagemodells für Aufgabenbearbeitungszeiten zeigten sich konzeptuelle und methodische Grenzen, besonders im Hinblick auf die (a) Messmethode, (b) Messgenauigkeit, (c) Generalisierbarkeit und (d) Ergebnisvalidierung.

Die Erhebung der Bearbeitungszeiten erfolgte durch die Testteilnehmerinnen und -teilnehmer, die die aktuelle Uhrzeit minutengenau vor und nach der Bearbeitung einer Aufgabe in ihr Testheft eintragen sollten. Hierdurch kann es zu Fehlern und/oder motivational bedingten fehlenden Werten (*missing by intention*) kommen. Die Missing-Rate in den Daten war allerdings gering und Plausibilitätschecks legten eine insgesamt hohe Datenkonsistenz nahe. Die Messung der Bearbeitungszeiten durch die Schülerinnen und Schüler hat die Datenqualität also höchstens geringfügig

beeinflusst.

Die potentiell geringere Messgenauigkeit durch minutengenaues anstatt höher auflösenden Messens stellt einen weiteren Kritikpunkt dar. Allerdings würden unpräzisere Messungen lediglich mehr „Noise“ in den untersuchten Zusammenhängen erzeugen. Dadurch würden die Zusammenhänge kleiner erscheinen als wenn die Messung akkurat wäre. In Anbetracht der hohen Varianzaufklärung von 94.3% ist es nicht unplausibel anzunehmen, dass die Messung eine hohe Präzision aufwies. Die berichteten Ergebnisse sind dennoch als untere Schranken zu interpretieren und könnten in Studien mit noch höherer Messpräzision noch deutlicher ausfallen.

Eine gute Alternative zu einer potentiell ungenaueren Messung im paper-&-pencil Kontext wäre, auf computerbasierte Messungen zurückzugreifen, die wesentlich hochauflösender (Millisekunden statt Minuten) sind. Jedoch sind Bearbeitungszeiten aus computerbasierten Messungen in erster Linie zur Erstellung von computerbasierten Tests geeignet. Deren Einsetzbarkeit für paper-&-pencil Tests ist unklar, da der Wechsel der Modalität die Reliabilität und Validität der Messung beeinflussen könnte. Obwohl Meta-Analysen zur Vergleichbarkeit von paper-&-pencil und computerbasierten Tests lediglich kleine *mode effects* berichten (z. B. Mead & Drasgow, 1993; Wang, Jiao, Young, Brooks & Olson, 2007, 2008), sollten drei Aspekte bei der Interpretation dieser Befunde beachtet werden: Erstens sind diese nur für Tests ohne Zeitlimit gültig. Bereits Mead und Drasgow (1993) zeigten, dass die fast perfekte Korrelation zwischen den Modalitäten für Power-Tests ohne Zeitlimit deutlich auf .72 für Tests mit Zeitlimit sank. Zweitens wurden in den Meta-Analysen die Mittelwertunterschiede untersucht. Mit Hinblick auf die Varianz-Kovarianz-Struktur sind aber durchaus Unterschiede zu erwarten (Schroeders & Wilhelm, 2011). Drittens sind Modalitäten-Effekte im Allgemeinen zwar klein; in spezifischen Situationen können diese aber auch wesentlich größer sein (van Lent, 2008). Ohne generalisierbares

Wissen über relevante Faktoren, die Modalitäten-Effekte erzeugen, sind Aussagen über die Adäquatheit von computerbasiert erhobenen und auf den paper-&-pencil Kontext übertragenen Bearbeitungszeiten nicht verlässlich treffbar.

Faktoren, die die Bearbeitungszeit in verschiedenen Modalitäten potentiell beeinflussen, könnten aus unterschiedlichen Anforderungen bezüglich Wahrnehmung und Motorik herrühren. Beispielsweise ließen sich Unterschiede im gemessenen Konstrukt auf das Scrollen langer Texte auf kleinen Bildschirmen mit geringer Auflösung (Bridgeman, Lennon & Jackenthal, 2003), auf das Klicken von Antwort-Buttons mit der Maus anstatt dem Ankreuzen des Antwortfeldes mit einem Stift (Pomplun, Frey & Becker, 2002) und auf das Verwenden der Tastatur anstatt des Aufschreibens auf Papier (Overton, Taylor, Zickar & Harms, 1996) zurückführen. Insgesamt ist also nicht auszuschließen, dass auch für Bearbeitungszeiten (und Zusammenhänge dieser mit anderen Variablen) Modalitäten-Effekte auftreten können. Deshalb ist es sicherer, Bearbeitungszeiten, die für die Konstruktion von paper-&-pencil Tests verwendet werden sollen, auch in paper-&-pencil Tests zu erheben und nicht auf computerbasiert erhobene Daten zurückzugreifen.

Die Generalisierbarkeit der Ergebnisse ist ebenfalls eingeschränkt. Diese sind nicht auf andere Populationen als deutschsprachige Neuntklässler generalisierbar. Besonders bei jüngeren Schülerinnen und Schülern könnte angenommen werden, dass diese wesentlich langsamer lesen und/oder schreiben. Ebenso könnten Wörter anderer Sprachen schneller oder langsamer zu lesen und/oder zu schreiben sein als deutsche Wörter.

Warum zeigte sich eine geringfügige Verschätzung in der Validierungsstichprobe? Diese kommt wahrscheinlich dadurch zustande, dass in der Original- und der Validierungsstichprobe Aufgaben unterschiedlicher Kompetenzbereiche verwendet wurden. Während in der Originalstichprobe, auf deren Grundlage die Entwicklung des

Vorhersagemodells stattfand, Aufgaben der Kompetenzbereiche *Fachwissen* und *Erkenntnisgewinnung* eingesetzt wurden, bestand die Validierungsstichprobe aus Aufgaben des Kompetenzbereiches *Bewertung*. Aufgaben, bei denen Bewertungen elaboriert werden müssen, scheinen demnach eine höhere Bearbeitungszeit zu benötigen als Aufgaben, die Fachwissen oder erkenntnistheoretisches Wissen abfragen. Eine visuelle Inspektion von Items mit dem Antwortformat *Erweiterte Antwort* ergab, dass Schülerinnen und Schüler tatsächlich etwas mehr bei Aufgaben aus dem Kompetenzbereich *Bewertung* schreiben. Das Auftreten eines solchen Effekts zeigt unbestritten auch einen Schwachpunkt der vorgeschlagenen Berechnungsformel, da das Antwortformat *Erweiterte Antwort* nur ein grober Proxy dafür ist, wie viel Text Schülerinnen und Schüler tatsächlich produzieren werden. Hier scheint abhängig vom Kompetenzbereich oder anderen (unbekannten) Einflussgrößen Variation zu existieren. Eine potenzielle Verbesserung des Vorhersagemodells könnte darin bestehen, die Größe des Antwortfeldes als Prädiktor miteinzubeziehen. Dazu müsste die Größe des Antwortfeldes je nach erwarteter Textmenge variiert werden. In der vorliegenden Studie war die Größe des Antwortfeldes jedoch über Aufgaben hinweg konstant.

4.4 Forschungsdesiderata

Die Desirata für Anschlussforschung ergeben sich direkt aus den drei Beiträgen und deren Grenzen. Trotzdem die Forschung zu Kontexteffekten auf eine lange Tradition zurückblickt, sind immer noch nicht alle Kontexteffekte und deren Implikationen hinreichend ergründet. Hierzu weitere Forschung zu betreiben ist insofern wichtig, als dass durch unbekannte Kontexteffekte unvorhersehbare negative Konsequenzen für die Güte der Messung auftreten können. Natürlich existieren auch Strategien für

die Handhabung unbekannter Kontexteffekte. Um das in der Einleitung beschriebene Beispiel der Federwaage weiter zu illustrieren: Wenn man nicht wüsste, wovon die unterschiedlichen Messwerte an verschiedenen Orten abhängen, könnte man einfach an vielen Orten messen und die Messwerte mitteln. So würde das unterschiedlich starke Wirken des Schwerefelds ausgemittelt. Analog könnten in Large-scale Assessments Messwiederholungen dazu dienen, das Wirken unbekannter Kontexteffekte auszumitteln. Diese Vorgehensweise ist aber aufgrund ihres hohen Aufwands problematisch. Die ganze Welt zu bereisen, um mit der Federwaage die Masse eines Objekts zu bestimmen, hat sicherlich einen gewissen touristischen Wert; mit Hinblick auf die Kosten kann dieses Vorgehen aber als suboptimal bewertet werden. Ähnlich verhält es sich im Large-scale Assessment, wo mehrfache Messungen zum Ausgleich von unbekannten Kontexteffekten aufgrund der hohen Kosten quasi ausgeschlossen sind. Weiterhin könnte das Testdesign auch prophylaktisch bezüglich aller potentiell möglichen Faktoren ohne Kenntnis über deren tatsächliche Wirkung balanciert werden. Dieses Vorgehen bietet sich aber ebenfalls nicht an, da hierdurch die Testheftanzahl schnell sehr groß werden würde, was höhere administrative Kosten zur Folge hätte. Es ist deshalb sinnvoller, relevante Kontexteffekte zu identifizieren, um diese dann geeigneter handhaben zu können.

Ein im Large-scale Assessment bisher wenig erforschter Kontexteffekt sind Carry-over-Effekte, also Effekte, die durch eine bestimmte Abfolge bestimmter Items oder Blöcke zustande kommen. So könnte ein Block, der aus Items desselben Kompetenzbereichs wie der vorherige Block besteht, wesentlich leichter zu bearbeiten sein als wenn vorher fachfremde Items zu bearbeiten gewesen wären. Weiterhin könnte ein sehr schweres Item die Frustration dermaßen steigern, dass das nächste Item nur sehr unmotiviert bearbeitet wird. Zu derartigen Effekten bedarf es weiterer Forschungsbemühungen.

Neben der Identifikation von Kontexteffekten, besitzt die Frage, warum diese auftreten, einen hohen Stellenwert, da so Erkenntnisse für die Optimierung der Messinstrumente gewonnen werden können. Einerseits können die Gründe für das Auftreten von Kontexteffekten in den Eigenschaften des Instruments gesucht werden. Dies war ein zentrales Anliegen dieser Dissertation. Neben den untersuchten Eigenschaften der Testheftschwierigkeit, der Anzahl an Items im Testheft und der Balancierung des Designs, sind natürlich noch viele weitere Eigenschaften denkbar. Zum Beispiel könnte ein Testheft, was aus Items verschiedener Kompetenzbereiche besteht, leichter oder schwerer sein als ein Testheft aus Items des gleichen Kompetenzbereiches.

Da Kontexteffekte immer erst bei der Messung – also bei der Interaktion von Testteilnehmern mit dem Messinstrument – entstehen, liegt es nahe, auch auf der Personenseite nach Determinanten für Kontexteffekte zu suchen. Bei der Untersuchung von Testhefteffekten wurde die Variation der Testheftschwierigkeit damit begründet, dass dadurch eine bessere Anpassung der Testheftschwierigkeit an die Fähigkeit von bestimmten Personengruppen erreicht werden kann. Die Annahme war, dass dies die Motivation steigern könnte, die Aufgaben sorgfältig zu bearbeiten. Allerdings waren Motivationseffekte nicht Gegenstand der Analysen. Demnach lauten die Forschungsfragen, die es zukünftig zu eruieren gilt: Kann durch eine bessere Passung zwischen Personenfähigkeit und Testheftschwierigkeit die Testteilnahmemotivation erhöht werden? Welcher funktionale Zusammenhang (keiner, linear, nicht-linear) besteht zwischen der Passung von Testschwierigkeit und Personenfähigkeit (objektive Unter- bzw. Überforderung) und der wahrgenommenen Unter- beziehungsweise Überforderung? Hat die Passung zwischen der Personenfähigkeit und Schwierigkeit des Tests einen Einfluss auf die Lösungswahrscheinlichkeit von Items? Technisch könnte diese Passung beispielsweise als Differenz des Personenparameters

und des Testheftparameters modelliert werden. Eine solche Modellierung ist im, in dieser Dissertation verwendeten, GLMM Framework allerdings nicht möglich. Alternativ könnte auf baysianische Verfahren zurückgegriffen werden, die eine wesentlich flexiblere Modellierung erlauben. Weiterhin wurde als statistisches Argument für die Optimierung der Passung die höhere Messpräzision angeführt. Auch dies sollte weiter untersucht werden: Wie verhält sich die Messpräzision in Abhängigkeit der Passung von Testheftschwierigkeit und Personenfähigkeit? Aus der Beantwortung dieser Forschungsfragen könnten weitere wichtige Erkenntnisse zur Optimierung von Messinstrumenten abgeleitet werden. Wenn es zum Beispiel gelänge, einen Test so zu gestalten, dass dieser sowohl für Gymnasiasten, Realschüler, Hauptschüler als auch für Förderschüler gleich motivierend ist und gleichzeitig auch noch präziser misst, wäre in Sachen Validität, Reliabilität und Testfairness viel gewonnen.

Ebenfalls interessant und lohnenswert wäre die Untersuchung der Dynamiken während der Testbearbeitung. Wie gestaltet sich der Zusammenhang zwischen Emotionen (Frustration, Langeweile, Spaß), der Motivation (Anstrengungsbereitschaft) und der Beanspruchung durch die Aufgaben? Hierbei könnten eine Vielzahl an Forschungsfragen beantwortet werden, beispielsweise: Steigt die Frustration in Abhängigkeit der Anzahl an falsch gelösten Aufgaben? Führt Frustration zu einer geringeren Anstrengungsbereitschaft? Steigt die Frustration bei inkorrekter Aufgabenbeantwortung in Abhängigkeit der persönlichen Relevanz, im Test gut abzuschneiden? Erhöht eine optimale Passung von Aufgabenschwierigkeit und Personenfähigkeit den Spaß an der Testbearbeitung? Ist man bei geringerer Motivation leichter frustriert? Zur Beantwortung solcher Fragen ist ein mikrolängsschnittliches Design, also eine Messung zu mehreren Zeitpunkten während der Testbearbeitung, notwendig. Außerdem sollte die Ableitung von konkreten Hypothesen auf einer starken theoretischen Basis beruhen, da eine explorative Ergründung aller potentiell möglichen Dyna-

miken und Interaktionen schnell sehr aufwändig würde. Beispielsweise könnte die *Control-Value Theory of Achievement Emotions* (Pekrun, 2006) einen fruchtbaren theoretischen Rahmen für solche Forschungsvorhaben bieten. Zur Bearbeitung von längsschnittlichen Daten müssen geeignete statistische Verfahren verwendet werden. Eine Auswahl solcher Verfahren werden von McArdle (2009) illustriert. Insbesondere das *Bivariate Dual Change Score Model* bietet sich an, um Dynamiken zweier zeitveränderlicher Variablen zu modellieren. Mit diesem Modell könnte beispielsweise untersucht werden, wie sich eine Veränderung der Motivation auf die Freude an der Testbearbeitung auswirkt und/oder ob eine Veränderung der Freude Auswirkungen auf die Motivation hat.

Mit dem vorgeschlagenen Vorhersagemodell lassen sich die für das Testdesign unverzichtbaren Aufgabenbearbeitungszeiten einfach und kostengünstig vorhersagen. Dieses Modell stellt allerdings nur einen Baustein bei der Bearbeitungszeitoptimierung von Testheften dar. Neben der Akkuratheit des Bearbeitungszeitkennwertes ist vor allem auch entscheidend, *welcher* Kennwert herangezogen wird. Im Einklang mit anderen Studien wurde der *Mittelwert* der Aufgaben vorhergesagt. Dies impliziert (bei Annahme einer Normalverteilung), dass 50% der Schülerinnen und Schüler die Aufgabe vollständig bearbeiten, die anderen 50% aber nicht mit der Bearbeitung fertig werden. Deshalb sollte für jeden konkreten Anwendungsfall sorgfältig überlegt und entschieden werden, ob der Mittelwert das optimale Kriterium ist. Alternativ wären andere Verteilungskennwerte wie zum Beispiel das .90-Quantil denkbar, was bedeuten würde, dass 90% der Schülerinnen und Schüler die Aufgabe schaffen.

Normalerweise wird im Testdesign allerdings auch nicht ein Komplettierungskriterium für die Aufgaben, sondern für die Testhefte festgelegt. So sollen zum Beispiel 90% der Schülerinnen und Schüler ihr Testheft vollständig bearbeiten. Bei der Zusammenstellung von Testheften aus Aufgaben ergibt sich aber folgendes – selten

diskutiertes – Problem: Zur Erreichung des Kriteriums auf Testheftebene kann nicht das gleiche Kriterium auf Aufgabenebene angesetzt werden (bei Korrelationen der Aufgabenbearbeitungszeiten unter eins – was quasi immer der Fall ist). Zum Beispiel ergibt die Summation der .90-Quantile der Bearbeitungszeiten der Aufgaben, die zu einem Testheft zusammengestellt werden sollen, nicht das .90-Quantil des Testhefts. Die Frage, die sich hier unmittelbar aufzwingt, ist, welches Quantil der Aufgaben verwendet werden muss, um das anvisierte Quantil auf Testheftebene zu treffen. Weiterhin könnte eine solche „Konversion“ auch von weiteren Testhefteigenschaften abhängen. Zum Beispiel wäre denkbar, dass das Aufgaben-Quantil mit zunehmender Anzahl an Aufgaben im Testheft kleiner wird.

4.5 Fazit

Systematik aus „Noise“ herauszuarbeiten ist eines der Hauptanliegen der meisten Wissenschaften. Hierbei spielen Messungen und Messinstrumente eine zentrale Rolle, indem durch diese definiert wird, was gemessen werden soll und wie die Messung vollzogen wird. In den meisten Disziplinen kommen Messinstrumente wiederholt und unverändert zum Einsatz. Hingegen werden die Messinstrumente im Large-scale Assessment in der Bildungsforschung immer wieder neu konstruiert und divergieren zusätzlich zwischen den Testteilnehmern. Hierdurch können Effekte, die auf die spezifische Konstruktion des Messinstruments zurückzuführen sind, die Messung und die Messergebnisse beeinflussen. Solchen Effekten widmete sich diese Dissertation.

Testhefte können durch ihre spezifische Zusammenstellung aus schwereren oder leichteren Items zusätzlich noch schwerer oder leichter sein. Dieser Effekt sollte im statistischen Auswertungsmodell berücksichtigt werden, um eine verzerrungsfreie Messung zu gewährleisten. Alternativ könnte die Testheftschwierigkeit aber auch

bereits bei der Konstruktion von Testheften konstant gehalten werden, um Testhefteffekte durch unterschiedliche Testheftschwierigkeiten zu vermeiden.

Auch die Positionierung von Items innerhalb von Testheften kann einen Einfluss auf die Messung ausüben. Solche Positionseffekte sind ein gut dokumentiertes Phänomen im Large-scale Assessment und konnten auch in den in dieser Dissertation analysierten Daten nachgewiesen werden. Zur Handhabung von Positionseffekten wird in vielen Large-scale Assessment Programmen die Strategie der Balancierung verwendet. Diese Strategie stellte sich als sinnvoll heraus. Häufig wird auch das Auftreten von Blockpaaren balanciert. Bezüglich dieser Vorgehensweise zeigte sich, dass diese keinen Einfluss auf die Messung hat. Somit kann der Blockpaarbalancierung im Testdesign eine niedrigere Priorität eingeräumt werden.

Weiterhin wurde im Rahmen dieser Dissertation ein Vorhersagemodell für Aufgabenbearbeitungszeiten auf Basis leicht verfügbarer Aufgabeneigenschaften vorgeschlagen, um einfach und kostengünstig zu diesen, für das Testdesign unverzichtbaren, Größen zu gelangen. Dieses Modell hilft sowohl Kosten zu sparen als auch die Validität der Messung zu garantieren. Alles in allem wurden also mehrere wichtige Beiträge zur Optimierung von Messinstrumenten im Large-scale Assessment geleistet.

Literatur

Albano, A. D. (2013). Multilevel modeling of item position effects. *Journal of Educational Measurement, 50*, 408-426.

Allen, N. L., Donoghue, J. R. & Schoeps, T. L. (2001). *The NAEP 1998 technical report* (No. NCES 2001-509). Washington, DC: National Center for Education Statistics. Retrieved from http://nces.ed.gov/nationsreportcard/pdf/main1998/2001509.pdf

Asseburg, R. & Frey, A. (2013). Too hard, too easy, or just right? The relationship between effort or boredom and ability-difficulty fit. *Psychological Test and Assessment Modeling, 55*, 92-104.

Barcikowski, R. S. (1972). A monte carlo study of item sampling (versus traditional sampling) for norm construction. *Journal of Educational Measurement, 9*, 209-214. doi:10.1111/j.1745-3984.1972.tb00954.x

Barcikowski, R. S. (1974). The effects of item discrimination on the standard errors of estimate associated with item-examinee sampling procedures. *Educational and Psychological Measurement, 34*, 231-237. doi:10.1177/001316447403400203

Bates, D. M. (2010). *lme4: Mixed-effects modeling with R.* Retrieved from http://lme4.rforge.r-project.org/book

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2014a). lme4: Linear mixed-effects models using Eigen and S4 (Version 1.1-6) [Computer software]. Retrieved from http://CRAN.R-project.org/package=lme4

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2014b). lme4: Linear mixed-effects models using Eigen and S4 (Version 1.1-7) [Computer software]. Retrieved from http://CRAN.R-project.org/package=lme4

Beaton, A. E. (1987). *Implementing the new design: The NAEP 1983-84 technical report* (No. NAEP-15-TR-20). Princeton, NJ: National Assessment of Educational Progress.

Beaton, A. E. (1988a). *Expanding the new design: The NAEP 1985-86 technical report* (No. NAEP-17-TR-20). Princeton, NJ: Educational Testing Service.

Beaton, A. E. (1988b). *The NAEP 1985-86 reading anomaly: A technical report.* Princeton, NJ: Educational Testing Service.

Beaton, A. E. & Zwick, R. (1990). *The effect of changes in the national assessment: Disentangling the NAEP 1985-86 reading anomaly* (Report No. ETS-17-TR-21). Princeton, NJ: Educational Testing Service.

Berger, V. F., Munz, D. C., Smouse, A. D. & Angelino, H. (1969). The effects of item difficulty sequencing and anxiety reaction type on aptitude test performance. *Journal of Psychology, 71*, 253-258. doi:10.1080/00223980.1969.10543091

Bergstrom, B. A., Gershon, R. C. & Lunz, M. E. (1994). *Computer adaptive testing: Exploring examinee response time using hierarchical linear modeling.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Nowick (Eds.), *Statistical theories of mental test scores* (pp. 395-479). Reading, MA: Addison-Wesley.

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443-459. 10.1007/BF02293801

Böhme, K., Leucht, M., Schipolowski, S., Porsch, R., Knigge, M. & Köller, O. (2010). Anlage und Durchführung des Ländervergleichs. In O. Köller, M. Knigge & B. Tesch (Hrsg.), *Sprachliche Kompetenzen im Ländervergleich* (pp. 65-85). Münster: Waxmann.

Brenner, M. H. (1964). Test difficulty, reliability, and discrimination as functions of item difficulty order. *Journal of Applied Psychology, 48*, 98-100. doi:10.1037/h0045738

Brennan, R. L. (1992). The context of context effects. *Applied Measurement in Education, 5*, 225-264. doi:10.1207/s15324818ame0503_4

Bridgeman, B., Lennon, M. L. & Jackenthal, A. (2003). Effects of screen size, screen resolution and display rate on computer-based test performance. *Applied Measurement in Education, 16*, 191-205.

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 221-256). Lanham, MD: Rowman & Littlefield Education.

Childs, R. & Jaciw, A. P. (2003). Matrix sampling of items in large-scale assessments. *Practical Assessment, Research & Evaluation, 8*(16). Retrieved from http://PAREonline.net/getvn.asp?v=8&n=16

Cizek, G. J. & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage Publications.

Cohen, J. E. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Cole, J. S., Bergin, D. A. & Whittaker, T. A. (2008). Predicting student achievement for low stakes tests with effort and task value. *Contemporary Educational Psychology, 33*, 609-624. doi:10.1016/j.cedpsych.2007.10.002

Cole, N. S. & Zieky, M. J. (2001). The new faces of fairness. *Journal of Educational Measurement, 38*, 369-382.

De Ayala, R. J. (2009). *The theory and practice of item-response theory*. New York: The Guilford Press.

De Boeck, P. & Wilson, M. (2004). A framework for item response models. In P. de Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 3-41). New York: Springer.

De Boeck, P. (2008). Random item IRT models. *Psychometrika*, 73, 533-559.

Debeer, D. & Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement, 50*, 164-185. doi:10.1111/jedm.12009

DeMars, C. (2002). Incomplete data and item parameter estimates under JMLE and MML estimation. *Applied Measurement in Education, 15*, 15-31.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological), 39*, 1-38.

Dorans, N. J., Pommerich, M. & Holland, P. W. (Eds.). (2007). *Linking and aligning scores and scales*. New York: Springer.

Eklöf, H. (2010). Skill and will: Test-taking motivation and assessment quality. *Assessment in Education: Principles, Policy & Practice, 17*, 345-356. doi:10.1080/0969594X.2010.516569

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Eysenck, H. J. & Eysenck, S. B. G. (1975). *Manual of the Eysenck Personality Questionnaire*. San Diego, CA: Educational and Industrial Testing Service.

Feldt, L. S. & Forsyth, R. A. (1974). An examination of the context effect in item sampling. *Journal of Educational Measurement, 11*, 73-82. doi:10.1111/j.1745-3984.1974.tb00975.x

Flaugher, R. L., Melton, R. S. & Myers, C. T. (1968). Item rearrangement under typical test conditions. *Educational and Psychological Measurement, 28*, 813-824. doi:10.1177/001316446802800310

Frey, A. & Bernhardt, R. (2012). On the importance of using balanced booklet designs in PISA. *Psychological Test and Assessment Modeling, 54*, 397-417.

Frey, A., Hartig, J. & Rupp, A. A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice, 28*(3), 39-53. doi:10.1111/j.1745-3992.2009.00154.x

Gelman, A., Meng, X. L. & Stern, H. S. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica, 6*, 733-807.

Giesbrecht, F. G. & Gumpertz, M. L. (2004). *Planning, construction, and statistical analysis of comparative experiments*. Hoboken, NJ: Wiley-Interscience.

Gonzalez, E. & Rutkowski, L. (2010). *Principles of multiple matrix booklet designs and parameter recovery in large-scale assessments*. In IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments (Vol. 3, pp. 125-156). Retrieved from http://www.ierinstitute.org/fileadmin/Documents/IERI_Monograph/IERI_Monograph_Volume_03_Chapter_6.pdf

Greene, W. H. (2011). *Econometric analysis* (7th ed.). Harlow, UK: Pearson Education Limited.

Gressard, R. P. & Loyd, B. H. (1991). A comparison of item sampling plans in the application of multiple matrix sampling. *Journal of Educational Measurement, 28*, 119-130.

Guttman, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society. Series B (Methodological), 29*, 83-100.

Haertel, E. (1989). *Report of the NAEP Technical Review Panel on the 1986 Reading Anomaly, the Accuracy of NAEP Trends, and Issues Raised by State-Level NAEP Comparisons* (No. CS-89-499). Washington, DC: US Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

Hahne, J. (2008). Analyzing position effects within reasoning items using the LLTM for structurally incomplete data. *Psychology Science Quarterly, 50*, 379-390.

Halkitis, P. N., Jones, J. P. & Pradhan, J. (1996). Estimating testing time: *The effects of item characteristics on response latency*. Paper presented at the Annual Meeting of the American Educational Research Association. New York, NY.

Hambleton, R. K. & Traub, R. E. (1974). The effects of item order on test performance and stress. *The Journal of Experimental Education, 43*, 40-46. doi:10.1080/00220973.1974.10806302

Hecht, M. (2014). eatDesign (Version 0.0.10) [Computer software]. Retrieved from http://RForge.R-project.org/projects/eat

Hecht, M., Roppelt, A. & Siegle, T. (2013). Testdesign und Auswertung des Ländervergleichs. In H. A. Pant, P. Stanat, U. Schroeders, A. Roppelt, T. Siegle & C. Pöhlmann (Hrsg.), *IQB-Ländervergleich 2012 – Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I* (S. 391-402). Münster: Waxmann.

Hohensinn, C., Kubinger, K. D., Reif, M., Holocher-Ertl, S., Khorramdel, L. & Frebort, M. (2008). Examining item-position effects in large-scale assessment using the Linear Logistic Test Model. *Psychology Science Quarterly, 50*, 391-402.

Hohensinn, C., Kubinger, K. D., Reif, M., Schleicher, E. & Khorramdel, L. (2011). Analysing item position effects due to test booklet design within large-scale assessment. *Educational Research and Evaluation, 17*, 497-509. doi:10.1080/13803611.2011.632668

Huck, S. W. & Bowers, N. D. (1972). Item difficulty level and sequence effects in multiple choice achievement tests. *Journal of Educational Measurement, 9*, 105-111. doi:10.1111/j.1745-3984.1972.tb00765.x

Hutcheson, G. D. & Sofroniou, N. (1999). *The multivariate social scientist: Introductory statistics using generalized linear models*. London: Sage Publications.

Johnson, C. M. & Lord, F. M. (1958). An empirical study of the stability of a group mean in relation to the distribution of test items among students. *Educational and Psychological Measurement, 18*, 325-329. doi:10.1177/001316445801800209

Joint Committee on Testing Practices. (2004). *Code of fair testing practices in education*. Washington, DC: American Psychological Association, Science Directorate. Retrieved from http://www.apa.org/science/programs/testing/fair-testing.pdf

Kiefer, T., Robitzsch, A. & Wu, M. (2014). TAM: Test Analysis Modules (Version 1.0-2.1) [Computer software]. Retrieved from http://CRAN.R-project.org/package=TAM

Kingston, N. M. & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement, 8*, 147-154. doi:10.1177/014662168400800202

Klosner, N. C. & Gellman, E. K. (1973). The effect of item arrangement on classroom test performance: Implications for content validity. *Educational and Psychological Measurement, 33*, 413-418. doi:10.1177/001316447303300224

Leary, L. F. & Dorans, N. J. (1985). Implications for altering the context in which test items appear: A historical perspective on an immediate concern. *Review of Educational Research, 55*, 387-413. doi:10.3102/00346543055003387

Lee, Y.-H. & Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modeling, 53*, 359-379.

Linacre, J. M. (2000). *Computer-adaptive testing: A methodology whose time has come* (MESA Memorandum No. 69). University of Chicago: MESA Psychometric Laboratory. Retrieved from http://www.rasch.org/memo69.pdf

Lord, F. M. (1962). Estimating norms by item-sampling. *Educational and Psychological Measurement, 22*, 259-267. doi:10.1177/001316446202200202

Lu, Y. & Sireci, S. G. (2007). Validity issues in test speededness. *Educational Measurement: Issues and Practice, 26*(4), 29-37. doi:10.1111/j.1745-3992.2007.00106.x

Marso, R. N. (1970). Test item arrangement, testing time, and performance. *Journal of Educational Measurement, 7*, 113-118. doi:10.1111/j.1745-3984.1970.tb00704.x

Mazzeo, J. & von Davier, M. (2014). Linking scales in international large-scale assessments. In L. Rutkowski, M. von Davier & D. Rutkowski (Eds.), Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis (pp. 229-258). Hoboken, NJ: Taylor and Francis.

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology, 60*, 577-605.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive-ability tests: A meta-analysis. *Psychological Bulletin, 114*, 449-458.

Meyers, J. L., Miller, G. E. & Way, W. D. (2009). Item position and item difficulty change in an IRT-based common item equating design. *Applied Measurement in Education, 22*, 38-60. doi:10.1080/08957340802558342

Mitzel, H. C., Lewis, D. M., Patz, R. J. & Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249-281). Mahwah, NJ: Erlbaum.

Mollenkopf, W. G. (1950). An experimental study of the effects on item-analysis data of changing item placement and test time limit. *Psychometrika, 15*, 291-315. doi:10.1007/BF02289044

Monk, J. J. & Stallings, W. M. (1970). Effects of item order on test scores. *Journal of Educational Research, 63*, 463-465.

Mood, C. (2010). Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European Sociological Review, 26*, 67-82. doi:10.1093/esr/jcp006

Moy, M. L. Y. (1973). *Item sampling: Optimum number of people and items.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, AERA, New Orleans, LA.

Munz, D. C. & Smouse, A. D. (1968). Interaction effects of item-difficulty sequence and achievement-anxiety reaction on academic performance. *Journal of Educational Psychology, 59*, 370-374. doi:10.1037/h0026224

Myerberg, N. J. (1975). *The effect of item stratification in multiple matrix sampling.* Paper presented at the Annual Meeting of the American Educational Research Association, Washington, DC.

Myerberg, N. J. (1979). The effect of item stratification on the estimation of the mean and variance of universe scores in multiple matrix sampling. *Educational and Psychological Measurement, 39*, 57-68. doi:10.1177/001316447903900109

Nachtigall, C., Kröhne, U., Enders, U. & Steyer, R. (2008). Causal effects and fair comparison: Considering the influence of context variables on student competencies. In J. Hartig, E. Klieme & D. Leutner (Eds.), *Assessment of competencies in educational contexts*. Cambridge, MA: Hogrefe & Huber Publishers.

Nakagawa, S. & Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution, 4*, 133-142. doi:10.1111/j.2041-210x.2012.00261.x

Neyman, J. & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica, 16*, 1-32. doi:10.2307/1914288

OECD. (2002). *PISA 2000 technical report*. Paris, France: Author.

OECD. (2005). *PISA 2003 technical report*. Paris, France: Author.

OECD. (2009). *PISA 2006 technical report*. Paris, France: Author.

OECD. (2012). *PISA 2009 technical report*. Paris, France: Author.

Overton, R. C., Taylor, L. R., Zickar, M. J. & Harms, H. J. (1996). The pen-based computer as an alternative platform for test administration. *Personnel Psychology, 49*, 455-464.

Owens, T. R. & Stufflebeam, D. L. (1969). An experimental comparison of item sampling and examinee sampling for estimating test norms. *Journal of Educational Measurement, 6*, 75-83. doi:10.1111/j.1745-3984.1969.tb00662.x

Pant, H. A., Stanat, P., Schroeders, U., Roppelt, A., Siegle, T. & Pöhlmann, C. (Hrsg.). (2013). *IQB-Ländervergleich 2012 – Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I.* Münster: Waxmann.

Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review, 18*, 315-341. doi:10.1007/s10648-006-9029-9

Plake, B. S. (1980). Item arrangement and knowledge of arrangement on test scores. *The Journal of Experimental Education, 49*, 56-58.

Plake, B. S., Ansorge, C. J., Parker, C. S. & Lowry, S. R. (1982). Effects of item arrangement, knowledge of arrangement, test anxiety, and sex on test performance. *Journal of Educational Measurement, 19*, 49-58. doi:10.1111/j.1745-3984.1982.tb00114.x

Plake, B. S., Patience, W. M. & Whitney, D. R. (1988). Differential item performance in mathematics achievement test items: Effect of item arrangement. *Educational and Psychological Measurement, 48*, 885-894. doi:10.1177/0013164488484003

Plake, B. S., Thompson, P. A. & Lowry, S. (1981). Effect of item arrangement, knowledge of arrangement, and test anxiety on two scoring methods. *The Journal of Experimental Education, 49*, 214-219.

Plummer, M. (2013). JAGS - Just another Gibbs sampler (Version 3.4.0) [Computer software].

Plummer, M. (2014). rjags: Bayesian graphical models using MCMC (Version 3.14) [Computer software]. Retrieved from http://CRAN.R-project.org/package=rjags

Plumlee, L. B. (1964). Estimating means and standard deviations from partial data–An empirical check on Lord's item sampling technique. *Educational and Psychological Measurement, 24*, 623-630. doi:10.1177/001316446402400316

Pomplun, M., Frey, S. & Becker, D. F. (2002). The score equivalence of paper-and-pencil and computerized versions of a speeded test of reading comprehension. *Educational and Psychological Measurement, 62*, 337-354.

Pugh, R. (1971). Empirical evidence on the application of Lord's sampling technique to Likert items. *The Journal of Experimental Education, 39*, 54-56.

R Core Team. (2014a). R: A language and environment for statistical computing (Version 3.1.0) [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

R Core Team. (2014b). R: A language and environment for statistical computing (Version 3.1.1) [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Robitzsch, A. (2009). Methodische Herausforderungen bei der Kalibrierung von Leistungstests. In A. Bremerich-Voss, D. Granzer & O. Köller (Hrsg.), Bildungsstandards Deutsch und Mathematik (S. 42-106). Weinheim: Beltz.

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion* (2., überarb. und erw. Aufl.). Bern: Verlag Hans Huber.

Rubin, D. B. (1984). Justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics, 12*, 1151-1172.

Sax, G. & Carr, A. (1962). An investigation of response sets on altered parallel forms. *Educational and Psychological Measurement, 22*, 371-376. doi:10.1177/001316446202200210

Sax, G. & Cromack, T. R. (1966). The effects of various forms of item arrangements on test performance. *Journal of Educational Measurement, 3*, 309-311. doi:10.1111/j.1745-3984.1966.tb00896.x

Scheetz, J. P. & Forsyth, R. (1977). *A comparison of simple random sampling versus stratification for allocating items to subtests in multiple matrix sampling.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York, NY.

Schnipke, D. L. & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In C. N. Mills, M. T. Potenza, J. J. Fremer & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments.* Mahwah, NJ: Lawrence Erlbaum Associates.

Schroeders, U. & Wilhelm, O. (2011). Equivalence of reading and listening comprehension across test media. *Educational and Psychological Measurement, 71*, 849-869. doi:10.1177/0013164410391468

Shoemaker, D. M. (1970a). Allocation of items and examinees in estimating a norm distribution by item sampling. *Journal of Educational Measurement, 7*, 123-128. doi:10.1111/j.1745-3984.1970.tb00706.x

Shoemaker, D. M. (1970b). Item-examinee sampling procedures and associated standard errors in estimating test parameters. *Journal of Educational Measurement, 7*, 255-262. doi:10.1111/j.1745-3984.1970.tb00726.x

Shoemaker, D. M. (1971a). *Principles and procedures of multiple matrix sampling* (Report No. SWRL-TR-34). Inglewood, CA: Southwest Regional Educational Lab.

Shoemaker, D. M. (1971b). An application of item-examinee sampling to scaling attitudes. *Journal of Educational Measurement, 8*, 279-282. doi:10.1111/j.1745-3984.1971.tb00938.x

Shoemaker, D. M. (1971c). Further results on the standard errors of estimate associated with item-examinee sampling procedures. *Journal of Educational Measurement, 8*, 215-220. doi:10.1111/j.1745-3984.1971.tb00928.x

Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a bayesian approach. *Journal of Educational Measurement, 42*, 375-394.

Sinharay, S. (2006). Bayesian item fit analysis for unidimensional item response theory models. *British Journal of Mathematical and Statistical Psychology, 59*, 429-449. doi:10.1348/000711005X66888

Sinaharay, S. & Johnson, M. S. (2003). *Simulation studies applying posterior predictive model checking for assessing fit of the common item response theory models* (Research Report No. RR-03-28). Princeton, NJ: Educational Testing Service.

Smouse, A. D. & Munz, D. C. (1968). The effects of anxiety and item difficulty sequence on achievement testing scores. *Journal of Psychology, 68*, 181-184. doi:10.1080/00223980.1968.10543421

Smouse, A. D. & Munz, D. C. (1969). Item difficulty sequencing and response style: A follow-up analysis. *Educational and Psychological Measurement, 29*, 469-472. doi:10.1177/001316446902900225

Stern, H. S. (2000). Asymptotic distribution of p values in composite null models: Comment. *Journal of the American Statistical Association, 95*, 1157-1160.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*, 677-680. doi:10.1126/science.103.2684.677

Steyer, R. & Eid, M. (2001). *Messen und Testen* (2., korr. Aufl.). Berlin: Springer.

Suppes, P. (1970). *A probabilistic theory of causality.* Amsterdam: North-Holland.

Swanson, D. B., Case, S. M., Ripkey, D. R., Clauser, B. E. & Holtman, M. C. (2001). Relationships among item characteristics, examinee characteristics, and response times on USMLE Step 1. *Academic Medicine, 76*, 114-116. doi:10.1097/00001888-200110001-00038

Towle, N. J. & Merrill, P. F. (1975). Effects of anxiety type and item difficulty sequencing on mathematics test performance. *Journal of Educational Measurement, 12*, 241-249. doi:10.1111/j.1745-3984.1975.tb01025.x

Tuerlinckx, F., Rijmen, F., Molenberghs, G., Verbeke, G., Briggs, D., van den Noortgate, W. & de Boeck, P. (2004). Estimation and software. In P. de Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 343-373). New York: Springer.

Van der Linden, W. J. & Glas, C. A. W. (2010). *Elements of Adaptive Testing.* New York: Springer.

Van der Linden, W. J., Veldkamp, B. P. & Carlson, J. E. (2004). Optimizing balanced incomplete block designs for educational assessments. *Applied Psychological Measurement, 28*, 317-331. doi:10.1177/0146621604264870

Van Lent, G. (2008). Important considerations in e-assessment. In F. Scheuermann & A. Guimarães Pereira (Eds.), *Towards a research agenda on computer-based assessment* (pp. 97-103). Ispra, Italy: European Commission, Institute for the Protection and Security of the Citizen.

Wainer, H. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Wang, S., Jiao, H., Young, M. J., Brooks, T. & Olson, J. (2007). A meta-analysis of testing mode effects in grade K-12 mathematics tests. *Educational and Psychological Measurement, 67*, 219-238.

Wang, S., Jiao, H., Young, M. J., Brooks, T. & Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K-12 reading assessments: A meta-analysis of testing mode effects. *Educational and Psychological Measurement, 68*, 5-24.

Weirich, S. (in Vorbereitung). *Kontexteffekte in Large-Scale Assessments* (Dissertation). Humboldt-Universität zu Berlin, Berlin.

Weirich, S., Haag, N. & Roppelt, A. (2012). Testdesign und Auswertung des Ländervergleichs: technische Grundlagen. In P. Stanat, H. A. Pant, K. Böhme & D. Richter (Hrsg.), *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik: Ergebnisse des IQB-Ländervergleichs 2011* (S. 277-290). Münster: Waxmann.

Weirich, S., Hecht, M. & Böhme, K. (2014). Modeling item position effects using generalized linear mixed models. *Applied Psychological Measurement, 38*, 535-548. doi:10.1177/0146621614534955

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6*, 473-492. doi:10.1177/0146621682006004080

Winship, C. & Mare, R. D. (1984). Regression models with ordinal variables. *American Sociological Review, 49*, 512-525. doi:10.2307/2095465

Whitely, S. E. & Dawis, R. V. (1976). The influence of test context on item difficulty. *Educational and Psychological Measurement, 36*, 329-337. doi:10.1177/001316447603600211

Wise, S. L. (2014). The utility of adaptive testing in addressing the problem of unmotivated examinees. *Journal of Computerized Adaptive Testing, 2*, 1-17.

Wise, S. L. & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*, 1-17. doi:10.1207/s15326977ea1001_1

Wolf, L. F. & Smith, J. K. (2005). The consequence of consequence: motivation, anxiety, and test performance. *Applied Measurement in Education, 8*, 227-242. doi:10.1207/s15324818ame0803_3

Wooldrigde, J. M. (2013). *Introductory econometrics: A modern approach* (5th ed.). Mason, OH: South-Western.

Wu, M. L., Adams, R. J., Wilson, M. R. & Haldane, S. A. (2007). ACER ConQuest version 2.0: Generalised item response modelling software [Computer software]. Camberwell, Victoria, Australia: ACER.

Yen, W. M. (1980). The extent, causes and importance of context effects on item parameters for two latent trait models. *Journal of Educational Measurement, 17*, 297-311. doi:10.1111/j.1745-3984.1980.tb00833.x

Yousfi, S. & Böhme, H. F. (2012). Principles and procedures of considering item sequence effects in the development of calibrated item pools: Conceptual analysis and empirical illustration. *Psychological Test and Assessment Modeling, 54*, 366-396.

Zwick, R. (1991). Effects of item order and context on estimation of NAEP reading proficiency. *Educational Measurement: Issues and Practice, 10*(3), 10-16. doi:10.1111/j.1745-3992.1991.tb00198.x

A model for the estimation of testlet response time in paper-and-pencil large-scale assessments

Hecht, M., Siegle, T. & Weirich, S. (eingereicht). A model for the estimation of testlet response time to optimize test assembly in paper-and-pencil large-scale assessments. Manuskript eingereicht zur Publikation in *Large-scale Assessments in Education*.

Abstract

Accurate response times are essential for assembling tests in educational large-scale assessment to ensure test validity and efficient testing. Because obtaining empirical response times in pilot studies is cost-intensive and also because this process is complicated in paper-and-pencil assessments, we propose a model-based approach for calculating response times from readily available testlet properties. This prediction formula was developed using the response time data of 334 high school students who worked on 93 testlets of a paper-and-pencil test measuring science achievement. A large proportion (94.3%) of the variance in response times was explained by number of items, number of words, and response type. Another sample of 1,386 students who worked on 125 additional science testlets was used to validate the initial findings. Overall, the proposed easy-to-use formula is suitable for providing accurate response times for test assembly at a low cost.

Keywords: response time, test assembly, large-scale assessment

A Model for the Estimation of Testlet Response Time to Optimize Test Assembly in Paper-and-Pencil Large-Scale Assessments

Multiple matrix sampling designs are the most commonly applied designs in educational large-scale assessments (Rutkowski, Gonzales, von Davier, & Zhou, 2014). The central idea of such designs is to construct several test forms—called *booklets* in paper-and-pencil tests—that are assembled from a large pool of testlets, which consist of a stimulus and one or several items. A major advantage of this approach is that each individual's workload can be held within acceptable limits while simultaneously covering a variety of different content domains across the test. One essential objective that needs to be fulfilled when compiling booklets is to ensure that the booklet can be reasonably completed within the prespecified testing time. Therefore, it is pivotal to know the testlet response times, which can be obtained in several ways. The most precise testlet response times would obviously be gained from direct measurement in a pilot study. However, this approach is usually laborious, time-consuming, and costly. Instead, testlet response times are often gauged by didactic experts in the process of testlet construction and development. However, the accuracy of and the consistency between experts' ratings might be—and often is—rather low. A promising alternative is to estimate response times from data that can be accessed without testing, for example, the number of words in a specific testlet. Although extensive amounts of research have addressed a variety of issues concerning response times in educational measurement in recent decades (for comprehensive literature reviews, see Lee & Chen, 2011; Schnipke & Scrams, 2002), surprisingly few studies have broached the idea of obtaining response time estimates from testlet (or item) properties. Halkitis, Jones, and Pradhan (1996) studied the degree to which item response time was related to item difficulty, item discrimination, and word count on a licensing examination. All of the predictors together accounted for half of the variance in the logs of item response time with word count as the strongest predictor

(R^2 = 27.2%), followed by item difficulty (R^2 = 16.2%), and item discrimination (R^2 = 6.8%). In the same vein, Bergstrom, Gershon, and Lunz (1994) identified item text length, (relative) item difficulty, item sequence, and position of the correct answer (in multiple-choice items) as relevant predictors. Furthermore, the presence of a figure had a strong impact on response times, although this might have been due to the administration of a separate illustration booklet. In data from a medical licensing examination, approximately 45% of the variance in item response time was explained by difficulty, the presence/absence of pictures, and the number of words (Swanson, Case, Ripkey, Clauser, & Holtman, 2001). The authors reported that "a logit change in item difficulty adds 14+ seconds," "the presence of a picture adds 12+ seconds," and "each word adds approximately 0.5 seconds" (p. 116). Even though empirical studies on this topic are rare, the results indicate that predicting response times from item properties is a worthwhile endeavor.

Test construction is not an end in and of itself but is always conducted with the goal of testing a specific population of students. Here, response times can provide valuable information about how to design the test as tests may function differently in different subpopulations. Consequently, this information is useful for tailoring tests to fit the needs of subpopulations with different time requirements. Research on the relations between person properties and response times is much more elaborate than research on item properties (again, see Lee & Chen, 2011; Schnipke & Scrams, 2002). However, the question of how student characteristics influence response times is typically addressed from a different angle with research that treats response time estimates as an auxiliary source of information for estimating individual ability (e.g., Wang & Hanson, 2005). Conversely, a person's ability is particularly important when studying response times. In a pioneering article on the estimation of response times (Thissen, 1983), the ability-latency relation was strongly moderated by the test content. Correlations between effective ability and "slowness" ranged from zero for a

spatial visualization test to .94 for a figural reasoning task. Analyses with contemporary statistical models (e.g., Klein Entink, Fox, & van der Linden, 2009) confirmed the complexity of this connection: Some studies found a negative relation between ability and speed, indicating that more capable test-takers spent more time on a task (Goldhammer & Klein Entink, 2011), whereas others reported the opposite result (Davison, Semmes, Huang, & Close, 2011). Besides the speededness of the measure, the relation has been further moderated by the anticipated consequences (low- vs. high-stakes testing) and personality traits such as conscientiousness or impulsivity (also see research on the "speed-accuracy tradeoff"). In summary, the relations between response times and student properties are not clear.

The popularity of research on response times has soared with the advent of the technology to measure them directly in computer-based assessments. Obviously, measuring response times in paper-and-pencil settings is far more complicated, and this is presumably the reason that almost all studies rely on computer-based data. However, data from computer-based tests may not be suitable for assembling paper-and-pencil tests because transposing the content from computer to paper may affect the reliability and validity of the measure. Although meta-analyses on the comparability of paper-based and computer-based assessments have reported only small to negligible cross-mode differences (e.g., Mead & Drasgow; 1993; Wang, Jiao, Young, Brooks, & Olson, 2007, 2008), three caveats must still be considered when interpreting such findings. First, this comparability holds only for unspeeded measures as Mead and Drasgow (1993) conclusively demonstrated that the almost perfect cross-mode correlation for timed power tests dropped considerably to .72 for speeded tests. Second, meta-analyses usually consider the mean structure but not the variance-covariance structure. Even if there are no mode effects for means, there might be mode-effects concerning the variances and covariances (Schroeders & Wilhelm, 2011). Third,

whereas cross-media differences are small in general, in a specific instantiation, substantial differences between test media may occur (van Lent, 2008)—and without generalizable knowledge about which factors affect the equivalence, it is difficult to determine the impact that a transition will have on response times. Factors affecting response times across media might consist of differences in the perceptual demands or the motor-skill requirement in the response procedure (Schroeders & Wilhelm, 2010). More precisely, differences across administration modes can result from scrolling down long texts on small screens with low screen resolution (Bridgeman, Lennon, & Jackenthal, 2003), clicking response buttons with a mouse instead of ticking the solution on a sheet of paper with a pen (Pomplun, Frey, & Becker, 2002), and using a keyboard instead of answering manually (Overton, Taylor, Zickar, & Harms, 1996). In summary, the change from paper to computer may alter the construct that the test administrator intends to measure. With this concern in mind, we decided not to assess response times on computers but to employ a paper-and-pencil assessment instead.

The Scope of the Present Research

The aim of the present study was to provide a well-founded and easy-to-use formula to calculate response times for testlets, stimuli, and items in order to optimize the assembly of paper-and-pencil tests in educational large-scale assessments. Furthermore, we explored whether response times depended on certain person properties in order to determine how to tailor test construction to the specific needs of specific subgroups of students. More precisely, we modeled response times as dependent on the testlet properties (a) number of items, (b) number of words, (c) response type (multiple-choice, short response, extended response), and (d) testlet difficulty. We simultaneously modeled them on the following student properties as well: (a) sex, (b) school track, and (c) competence. In a second step, we validated this empirically obtained model in a new sample of testlets and students.

Method

Participants

Study 1 was used to develop the prediction formula. The sample consisted of 334 students in Grade 9 (49.4% girls, 2.1% did not indicate their sex) with an average age of 15.5 years (*SD* = 0.75) from four academic-track schools (56.9%) and three intermediate-track schools (43.1%). Academic-track schools prepare students for university enrollment, whereas students in intermediate-track schools often pursue a vocational education. Participation was voluntary, and students were not rewarded or graded in any way. Data were collected in the spring of 2010.

Study 2 was conducted for validation purposes. The data collection took place in the fall of 2010. All 1,386 students were 10th graders from intermediate-track schools, and almost half of them were girls (48.0%; 1.9% did not indicate their sex).

Design and Procedure

In Study 1, we distributed 93 testlets, each of which contained a stimulus and one to five (*M* = 1.86) items. They originated from a large pool of testlets that were designed to measure the *German Science Education Standards* (for details on the development and evaluation of these standards, see Kremer et al., 2012; Neumann, Fischer, & Kauertz, 2010; Pant et al., 2013). The conceptual core of educational standards is very similar to the idea of *scientific literacy* (e.g., Holbrook & Rannikmae, 2009) and contains four subdomains: content knowledge, scientific inquiry, decision making, and communication. Because the test development for these subdomains was time-delayed, only testlets measuring content knowledge and scientific inquiry were available in Study 1. The required response types consisted of either choosing an answer (multiple-choice), writing one or several words (short response), or writing several sentences (extended response). Figure 1 displays an example testlet from the subdomain *scientific inquiry* consisting of a stimulus and a multiple-choice

item. We employed an incomplete block design (e.g., Frey, Hartig, & Rupp, 2009) with 24 booklets that were randomly administered to the students. The test construction process began by grouping testlets into clusters of 20 min. Eight clusters were assembled for each science subject (biology, chemistry, physics). In the next step, each of these clusters was assigned to two booklets. Following this procedure, each booklet contained three clusters, one for each science domain. Because the unspeeded response time was the variable of interest, all students were provided with sufficient time to complete the test. Before and after working on each testlet, students were asked to record the time in the test booklet. In order to standardize the time recording, a clock was positioned in front of the class.

In Study 2, the design and procedure were similar to Study 1, except for three changes: (a) the topic of the testlets consisted of another scientific competence, decision making, (b) booklets contained two clusters of 20 min length equaling 40 min of total testing time, and (c) booklets contained either one cluster of chemistry and one cluster of physics or two clusters of biology. A total of 51 booklets were assembled and randomly administered to the students.

Statistical Analyses

We specified five consecutive *linear mixed models* (LMM) to predict the response time using the characteristics of testlets and students. Response times were recorded in seconds—y_{jt} was the time student j worked on testlet t. As a consequence of the multiple matrix sampling design, students and testlets were partially crossed, that is, not all combinations of j and t were observed. Because booklets were distributed to students randomly, missingness was also *completely at random* (MCAR). LMM software such as the R (R Core Team, 2014) package *lme4* (Bates, Mächler, Bolker, & Walker, 2014) is able to handle MCAR adequately.

The first model in the series contained only an intercept α_0, a student parameter, θ_j, a testlet parameter, β_t, and a Student × Testlet interaction parameter ε_{jt}:

$$y_{jt} = \alpha_0 + \theta_j + \beta_t + \varepsilon_{jt} \quad (1)$$

In this model, the intercept α_0 is the overall mean testlet response time, θ_j is the deviation of student *j* from this mean, and β_t is the deviation of testlet *t*. The term ε_{jt} is the interaction of a specific student *j* with a specific testlet *t*. As the purpose of the present study was to investigate the effects of testlet and person properties on the response time, the point estimates for students and testlets were of less interest. Thus, students, testlets, and interactions were each modeled as *random effects*, assuming a normal distribution with means of zero and variances of σ_θ^2, σ_β^2, and σ_ε^2. These variances indicated the extent to which the students and testlets diverged from the overall mean on average. The testlet and student properties in subsequent models were expected to explain this variability in response time.

In the second model, the number of items (N_{items}) and the total number of words in the testlets (N_{words}) were added to the model as predictors (*fixed effects*):

$$y_{jt} = \alpha_0 + \gamma_{items} N_{items} + \gamma_{words} N_{words} + \theta_j + \beta_t + \varepsilon_{jt} \quad (2)$$

where the intercept α_0 is the mean response time of the (hypothetical) testlets with zero items and zero words. The variance of testlets, σ_β^2, is conditional on the effects of testlet properties and could be interpreted as the remaining unexplained variance. The main purpose of Model 2 was to estimate the additional time that a single item added to the response time needed to complete a testlet. The estimate of this effect, γ_{items}, was used as a *fixated* effect in the following models to facilitate interpretation and to avoid estimation problems due to collinearity effects caused by high correlations between the number of items and the remaining predictors. Note that the term fixated is used for fixed effects that are fixed to a specific value.

Model 3 added the centered numbers of multiple-choice items (N_{MC}), short response items (N_{SR}), and extended response items (N_{ER}) and the centered difficulty of the testlet (X_{diff}):

$$y_{jt} = \alpha_0 + \gamma^*_{items} N_{items} + \gamma_{words} N_{words} + \gamma_{MC} N_{MC} + \gamma_{SR} N_{SR} + \gamma_{ER} N_{ER} + \gamma_{diff} X_{diff} + \theta_j + \beta_t + \varepsilon_{jt} \quad (3)$$

The centering of the response type variables offered a simple interpretation: Whereas the effect of number of items indicates how much response time is needed for an item in general, the response type effects express the additional time needed for items of a particular response type. The testlet difficulty, X_{diff}, was estimated using a partial credit model with the software package ConQuest 2.0 (Wu, Adams, Wilson, & Haldane, 2007) and centered afterwards. Thus, if the difficulty changed by one logit, the response time increased by a value of γ_{diff}.

In Model 4, student properties were added to explain the variability in students' response times. The predictors were sex (Z_{sex}), school track (Z_{track}), and students' competence (Z_{comp}). We used the Greek letter δ instead of γ to distinguish between the effects of student properties and testlet properties:

$$\begin{aligned} y_{jt} = {} & \alpha_0 + \delta_{sex} Z_{sex} + \delta_{track} Z_{track} + \delta_{comp} Z_{comp} + \gamma^*_{items} N_{items} + \gamma_{words} N_{words} \\ & + \gamma_{MC} N_{MC} + \gamma_{SR} N_{SR} + \gamma_{ER} N_{ER} + \gamma_{diff} X_{diff} + \theta_j + \beta_t + \varepsilon_{jt} \end{aligned} \quad (4)$$

The variability in students was modeled as conditional on the effects of student properties and represented the unexplained variability in students' response times. Students' competence estimates (centered WLEs) came from the same item response model as the testlet difficulties. The dichotomous variables sex (boys vs. girls) and school track (intermediate vs. academic) were effect coded. Instead of using the default effect codes (i.e., -1 and 1), we modified them according to the proportions of the respective groups in our sample (this is essentially equivalent to centering the effect codes). For the variable school track, the centered effect codes were $Z_{track1} = -1.14$ for the intermediate track and $Z_{track2} = 0.86$ for the academic track. Because the sample contained almost equal proportions of boys and

girls, the centered effect codes for the variable sex were Z_{sex1} = -1.01 for boys and Z_{sex2} = 0.99 for girls. The advantage of centered effect codes is that the effect is estimated for equal proportions of the two groups (50%) even though the distributions in the sample may be different. As a consequence, the intercept does not change when such centered effect coded variables are entered into the model. Furthermore, we included all significant Student × Testlet interactions in the model.

For the final model, Model 5, all nonsubstantial effects from Model 4 were excluded to derive a prediction formula that could be easily implemented to calculate the response times for testlets and items on paper-and-pencil tests in large-scale assessments.

All models were estimated with the function *lmer* from the *lme4* package (Bates et al., 2014). Confidence intervals were bootstrapped using the lme4-function *bootMer* with 10,000 simulations and the function *boot.ci* from the package *boot* (Canty & Ripley, 2014). An estimate is considered significantly different from zero if zero is outside the 95% confidence interval. The main reason for using the bootstrapped and therefore potentially asymmetric confidence intervals is that *lme4* does not provide standard errors for variance estimates because "in most cases summarizing the precision of a variance component estimate by giving an approximate standard error is woefully inadequate" (Bates, 2010, p. 19).

The log of the response times is often reported as log-normally distributed and analyzed accordingly. For the data at hand, the log of the response times did not better approximate a normal distribution. Thus, the original response time estimates were used. The approximate normal distribution of item response times in our data may be due to the fact that the response times for any items that were skipped were treated as missing values. In computer-based assessment, however, such items would have skewed the response time distribution because they would have had near-zero response times.

Results

Descriptive Statistics

Table 1 shows the descriptive statistics for the empirical testlet response-time and testlet properties in Study 1 (for prediction) and Study 2 (for validation). The main difference between the testlets used in these two studies was their length. The Study 2 testlets, which measured decision-making competence in science, contained more items (M_1 = 1.86 vs. M_2 = 2.92) and nearly twice as many words (M_1 = 175.14 vs. M_2 = 322.82) than the Study 1 testlets, which measured content knowledge and scientific inquiry skills. These differences in testlet length were reflected in a much higher average response time in Study 2 (M_1 = 148.20 vs. M_2 = 277.66). The correlations between the empirical testlet times and testlet properties are shown in Table 2. Not surprisingly, the number of items was highly correlated with response time (r_1 = .87, r_2 = .70). The correlations between the number of words and response time were also large (r_1 = .75, r_2 = .64). Further, the number of items and the number of words were also substantially correlated (r_1 = .68, r_2 = .65). The correlations between the response type variables and the number of words ranged from .32 to .50 in Study 1.

Because of the high intercorrelations between the number of items and the other variables, the effect was first estimated (in Model 2) and fixated in all subsequent models. This approach allowed us to disentangle the impact of number of items and the other variables on response time despite the high correlations. To provide a better understanding of the relations between variables, partial correlations (i.e., correlations controlled for the number of items) are reported in the upper triangle of Table 2 (calculated with the R package *parcor*; Krämer & Schäfer, 2014). For example, as the number of multiple-choice items increased—relative to items with other response types—the testlet response time decreased in both studies (r_1 = -.43, r_2 = -.58). By contrast, as the number of extended response items increased, the testlet response time was higher (r_1 = .27, r_2 = .48). The relation between testlet

difficulty and response type was as follows: Testlets that contained more multiple-choice items were easier (r_1 = -.24, r_2 = -.41), whereas those that contained more extended response items were associated with greater testlet difficulty (r_1 = .35, r_2 = .45).

Models

Table 3 displays the results of all five consecutive models. In Model 1, the intercept representing the overall average testlet response time was α_0 = 149.4 s. The deviations of testlets (SD_β = 65.1, 95% CI [55.4, 75.4]) and students (vs. SD_θ = 27.4, 95% CI [24.0, 30.9]) were significantly different from zero—thus, there was indeed a substantially large amount of variability that could be explained by the properties of testlets and students in further models. The purpose of Model 2 was to estimate the effect of the number of items so that the parameter could be included as a fixated effect in subsequent analyses. This effect amounted to γ_{items} = 43.8 s, which means that adding one item to the testlet increased the response time by 43.8 s on average. The two testlet properties in Model 2, number of items and number of words, explained 83.4% of the variability in testlet response time. Further, Model 2 possessed a smaller BIC and AIC than Model 1 (see Table 3), indicating that this model was more suitable for explaining the data at hand. In Model 3, the centered number of items of any response type (i.e., multiple-choice, short response, or extended response) and the centered testlet difficulty were added as predictors. The effect of the *multiple-choice* response type was estimated as γ_{MC} = -24.4 s, that is, students were able to answer multiple-choice items faster than the overall average. In other words, adding one multiple-choice item increased the response time by $\gamma_{items} - \gamma_{MC}$ = 43.8 - 24.4 = 19.4 s. The effects for short response and extended response items were γ_{SR} = 2.8 and γ_{ER} = 14.5, respectively. Comparing a multiple-choice item to an extended response item (with an equal number of words and difficulty) yielded a difference of 14.5 - (-24.4) = 38.9 s because writing a paragraph takes much longer than just ticking boxes. Surprisingly, the difficulty of the task played no role as indicated by

the near-zero and nonsignificant effect γ_{diff} = 1.1, 95% CI [-3.3, 5.4]. The intercept was also near zero (α_0 = 2.6, 95% CI [-11.4, 16.2]) because a hypothetical testlet with zero words and zero items would take no time to complete. Further, a (hypothetical) testlet with just a single stimulus but no items had a response time that depended on only the words of the stimulus that needed to be read. For every word in the testlet, the response time increased by γ_{words} = 0.37 s. Thus, increasing the text length by 100 words would add 37 s to the predicted response time.

Besides the testlet properties, Model 4 additionally included student properties. All student properties that were considered in our study—sex, school track, and students' competence—exhibited only very marginal nonsignificant effects (δ_{sex} = 0.91, 95% CI [-3.0, 4.7], δ_{track} = 1.6, 95% CI [-2.5, 5.8], and δ_{comp} = 0.18, 95% CI [-4.4, 4.7]). Therefore, it did not seem necessary to adjust the response times for booklets that were specifically designed for these subsamples (boys vs. girls, intermediate vs. academic track, more vs. less competent students). However, two interactions between testlet and person properties were relevant and therefore included in Model 4, that is, Sex × Extended Response (6.1, 95% CI [1.6, 10.6]) and School Track × Extended Response (16.6, 95% CI [12.2, 21.1]). Girls worked 6.1 s longer on extended response items. Concerning school track, students who were enrolled in an academic-track school worked 16.6 s longer on an item with an extended response format than students enrolled in an intermediate-track school. These differences should be taken into account when tests contain a sufficient number of extended response items and need to be tailored to these subgroups.

For Model 5, all of the nonsubstantial predictors from Model 4 were excluded to derive a formula that would be easy to use to calculate response time estimates. Although the effect of short responses was nonsignificant, it was retained in the prediction model in order

to facilitate confusion-free handling. This final and best fitting model explained 94.3% of the variability in testlets:

$$\hat{y} = 43.8N_{items} + 0.39N_{words} - 25.9N_{MC} + 2.2N_{SR} + 14.7N_{ER} + 6.1Z_{sex}N_{ER} + 16.6Z_{track}N_{ER} \qquad (5)$$

Examples

We will now present examples that show how to use this formula, which can be applied to calculate response times for (a) stimuli, (b) items, and (c) testlets. In the simplest case of a stimulus, all predictors are set to zero except for the number of words. The response time for a stimulus with $N_{\text{words}} = 100$ is then $y_{(a)} = 0.39 * 100 = 39$ s. An extended response item with $N_{\text{words}} = 100$ will take $y_{(b)} = 43.8 * 1 + 0.39 * 100 + 14.7 * 1 = 97.5$ s to complete. If this item will be employed in academic-track schools, 16.6 s should be added: $y_{(b)2} = y_{(b)} + 16.6 * Z_{\text{track}=2} * N_{\text{ER}} = 97.5 + 16.6 * 1 * 1 = 114.1$ s. For intermediate-track schools, 16.6 s should be subtracted: $y_{(b)1} = y_{(b)} + 16.6 * Z_{\text{track}=1} * N_{\text{ER}} = 97.5 + 16.6 * (-1) * 1 = 80.9$ s. For a testlet, two approaches are feasible: either applying the formula to the entire testlet or summing the response times of the elements. We combined the previously used stimulus and two of the previously used extended response items into a testlet. When applying the formula, this yielded: $y_{(c)1} = 43.8 * 2 + 0.39 * (100 + 100 + 100) + 14.7 * 2 = 234$ s. Alternatively, the separately calculated response times can be summed across the three elements: $y_{(c)2} = y_{(a)} + 2 * y_{(b)} = 39 + 2 * 97.5 = 234$ s. An asset of this formula is that it allows various testlets to be assembled from items with precalculated response time without the need to apply the formula to the testlet.

Validation

The prediction of testlet response times using the testlet properties from Equation 5 is of course very accurate for the original sample because the formula was derived from this sample. The mean predicted response time of the 93 testlets was $M_1 = 149.78$ ($SD_1 = 64.02$). The predicted value deviated only a trivial amount from the empirical mean of $M_{1\text{diff}} = 1.58$

(SD_{1diff} = 20.06). In the validation sample with 125 additional testlets, the mean of the predicted response times was M_2 = 253.79 (SD_2 = 99.31)—a deviation from the empirical mean of M_{2diff} = -23.87 (SD_{2diff} = 49.88)—thus resulting in an underestimation of 8.6%.

Discussion

Accurate response times of testlets and items are crucial for assembling the booklets of a paper-and-pencil test that will be used in large-scale assessments. The most accurate response times can certainly be obtained by pilot testing the testlets and items, but this process is time-consuming and expensive. Nevertheless, even for pilot testing, it is necessary to have some initial response time estimates for testlets. Obtaining response time estimates from available testlet and item properties is a quick, convenient, and low-cost alternative to extensive pilot testing or expert ratings. To derive an empirically based formula, we collected response times from a sample of high school students who worked on science testlets. On the basis of these empirical data, we acquired a sound prediction model (Model 5) that can be used to estimate response times for stimuli, items, and testlets from (a) the number of items, (b) the number of words, and (c) the response type. These are all easy-to-obtain properties that are available without cost-intensive pilot testing.

Our results are plausible and in line with previous research. Number of words was also identified as a relevant predictor in the studies by Halkitis et al. (1996), Bergstrom et al. (1994), and Swanson et al. (2001). In our data, it took 0.39 s to process one word, a value that is close to Swanson's estimate of (approximately) 0.5 s. In our assessment of student properties, we found no main effects of sex or school track. These findings are consistent with Bergstrom's findings, which suggested that "examinee characteristics are generally not related to response time" (p. 13). For test designers, this is a satisfactory outcome because there is no need to construct separate test forms, for example, for academic-track and intermediate-track schools.

The statistical method (linear mixed models) that we employed allowed us to estimate Item × Student interactions. We investigated these in an exploratory fashion and found that girls and academic-track students invested more time in writing extended responses. Test designers selecting test items should be aware of such additional influences on response time. Booklets with a disproportionately high number of such items may differentially affect the response times of students in specific educational tracks. However, we would like to suggest that test administrators carefully consider the political implications of allocating different times to subgroups because questions of test fairness may emerge. Furthermore, student estimates from studies with different time restrictions may be difficult to compare. However, using the final prediction formula to assemble and optimize test booklets that are administered to all students from a certain population should not be problematic.

To validate our empirically derived prediction formula, a second sample of students worked on a different set of science testlets. Applying Equation 5 and comparing predicted and empirical response times yielded an average underestimation of 8.6%. This prediction bias is probably due to the different competences measured in the two studies. Whereas testlets that measured content knowledge and scientific inquiry were used in the original sample, in the validation study (Study 2), the competence in question was decision making in science. Such items require test-takers to thoroughly elaborate on a decision or an evaluation, a process that appears to be more time-consuming than answering the Study 1 items, which assessed knowledge about science and scientific procedures. An inspection of students' responses to the extended response items suggested that students indeed wrote more when they answered items on competence in decision making. Such an effect of the content domain is undeniably one of the major threats to the chosen prediction model because the response type extended response is just a rough proxy for the actual amount of text that is produced. Depending on the competence that is measured or on other (unknown) variables, there might

be nontrivial variation in the amount of text students produce. Furthermore, the results of the present study may not generalize to populations other than German-speaking students or students in Grade 9. Sentences in other languages might be either more concise or lengthier and thus faster or slower to read and write. Furthermore, younger (e.g., primary school) students might be much slower at reading the same amount of text. A further limitation was the use of response times from paper-and-pencil assessments because such measurements are less precise in comparison with a computer-based assessment. On the other hand, using computer-based response times to construct paper-and-pencil tests is also not a feasible option as mode effects may jeopardize the validity of the measurement and lead to severe biases. More research is needed to investigate and predict mode effects on response times and to describe their implications. Further, potentially less precise paper-and-pencil response times would just add "noise" to the relations under investigation. Thus, relations would appear smaller than they actually were if response time measurement was accurate. Given the large amount of explained variance (94.3%), it is reasonable to assume that our measurements were quite accurate. Still, the reported results are lower boundaries and may even be more pronounced in other studies with even higher measurement precision.

In this study, we used a modeling approach that explicitly allowed us to divide the response time variance into variance accounted for by items and variance accounted for by persons. The presented mixed models offer three methodological advantages over the often-used standard regression analyses. First, mixed models enabled us to consider Item × Student interactions that could provide additional information for test construction. Second, mixed models can adequately account for the data structure in large-scale assessments where students work on different item subsets assembled in various booklets. Third, mixed models offered higher test power because the response time data were not aggregated (and thus not

reduced) across persons or items. In other words, more data points were available for the estimation of model parameters.

Another important methodological issue was the criterion that was targeted in our prediction model. In line with other studies, we predicted the mean response time. This implies that 50% of the students will complete the testlet in this amount of time, and 50% will not (under the assumption of a normal distribution). Test administrators should consider whether the mean response time is the correct choice for the specific application of the test. One may argue that some other criterion will offer a worthwhile alternative. For instance, it may be reasonable to enable 90% of the students to complete the testlet, in which case the .90 quantile of the response time distribution would be chosen. A related—and rarely discussed—issue is the aggregation of item or testlet response time into booklet response time. Here, the standard approach is to sum the response times across items or testlets to derive the booklet response time, which equals the time that is available per student. This technique might not work for all criteria because students' rank ordering of response times will change from testlet to testlet if their correlations are below 1 (which is usually the case). This implies that testlet time cannot simply be added to calculate booklet time if certain criteria are used. For instance, if 90% of the students are expected to complete their booklets, it is not correct to sum the .90 quantiles of the testlets that comprise the booklet. Instead, some lower quantile would be the right choice in this case. Further research is needed to identify the testlet quantiles that lead to a certain target quantile at the booklet level.

To conclude, the present study provides an empirically derived formula for the prediction of response times for items, stimuli, and testlets for paper-and-pencil tests. Although the prediction was not perfect, its simplicity and cost-efficiency compensates for the minor inaccuracies that we identified. Besides, response times are indispensable for the construction of test instruments in large-scale assessments and have to be gauged somehow.

Our prediction formula offers a convenient way and might even outperform other methods such as expert ratings.

References

Bates, D. M. (2010). *lme4: Mixed-effects modeling with R*. Retrieved from http://lme4.r-forge.r-project.org/book/

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4 (Version 1.1-7)*. Retrieved from http://CRAN.R-project.org/package=lme4

Bergstrom, B. A., Gershon, R. C., & Lunz, M. E. (1994). *Computer adaptive testing: Exploring examinee response time using hierarchical linear modeling*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2003). Effects of screen size, screen resolution and display rate on computer-based test performance. *Applied Measurement in Education*, *16*, 191–205.

Canty, A., & Ripley, B. (2014). *boot: Bootstrap R (S-Plus) functions (Version 1.3-13)*. Retrieved from http://CRAN.R-project.org/package=boot

Davison, M. L., Semmes, R., Huang, L., & Close, C. N. (2011). On the reliability and validity of a numerical reasoning speed dimension derived from response times collected in computerized testing. *Educational and Psychological Measurement*, *72*, 245–263.

Frey, A., Hartig, J., & Rupp, A. A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice*, *28*(3), 39–53.

Goldhammer, F. & Klein Entink, R. H. (2011). Speed of reasoning and its relation to reasoning ability. *Intelligence*, *39*, 108–119.

Halkitis, P. N., Jones, J. P., & Pradhan, J. (1996). *Estimating testing time: The effects of item characteristics on response latency.* Paper presented at the Annual Meeting of the American Educational Research Association. New York, NY.

Holbrook, J., & Rannikmae, M. (2009). The meaning of scientific literacy. *International Journal of Environmental and Science Education*, *4*, 275–288.

Krämer, N., & Schäfer, J. (2014). *parcor: Regularized estimation of partial correlation matrices (Version 0.2-6)*. Retrieved from http://CRAN.R-project.org/package=parcor

Kremer, K., Fischer, H. E., Kauertz, A., Mayer, J., Sumfleth, E., & Walpuski, M. (2012). Assessment of standard-based learning outcomes in science education: Perspectives from the german project ESNAS. In S. Bernholt, K. Neumann, & P. Nentwig (Eds.), *Making it tangible: learning outcomes in science education* (pp. 201–218). Münster: Waxmann.

Klein Entink, R. H., Fox, J. P., & van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika*, *74*, 21–48.

Lee, Y.-H., & Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modeling*, *53*, 359–379.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive-ability tests: A meta-analysis. *Psychological Bulletin*, *114*, 449–458.

Neumann, K., Fischer, H. E., & Kauertz, A. (2010). From PISA to educational standards: The impact of large-scale assessments on science education in Germany. *International Journal of Science and Mathematics Education*, *8*, 545–563.

Overton, R. C., Taylor, L. R., Zickar, M. J., & Harms, H. J. (1996). The pen-based computer as an alternative platform for test administration. *Personnel Psychology*, *49*, 455–464.

Pant, H. A., Stanat, P., Schroeders, U., Roppelt, A., Siegle, T., & Pöhlmann, C. (2013). *The IQB national assessment study 2012 – competencies in mathematics and the sciences at the end of secondary level*. Münster: Waxmann.

Pomplun, M., Frey, S., & Becker, D. F. (2002). The score equivalence of paper-and-pencil and computerized versions of a speeded test of reading comprehension. *Educational and Psychological Measurement*, *62*, 337–354.

R Core Team. (2014). *R: A language and environment for statistical computing. (Version 3.1.1)*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Rutkowski, L., Gonzales, E., von Davier, M., & Zhou, Y. (2014). Assessment design for international large-scale assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski, D. (Eds.), *Handbook of international large-scale assessment: background, technical issues, and methods of data analysis* (pp. 75–95). Boca Raton: CRC Press.

Schnipke, D. L., & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Schroeders, U., & Wilhelm, O. (2010). Testing reasoning ability with handheld computers, notebooks, and paper and pencil. *European Journal of Psychological Assessment, 26*, 284–292.

Schroeders, U., & Wilhelm, O. (2011). Equivalence of reading and listening comprehension across test media. *Educational and Psychological Measurement*, *71*, 849–869.

Swanson, D. B., Case, S. M., Ripkey, D. R., Clauser, B. E., & Holtman, M. C. (2001). Relationships among item characteristics, examinee characteristics, and response times on USMLE Step 1. *Academic Medicine*, *76*, 114–116.

Thissen, D. (1983). Timed testing: An approach using item response theory. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testin*g (pp. 179–203). New York, NY: Academic Press, Inc.

van Lent, G. (2008). Important considerations in e-assessment. In F. Scheuermann & A. Guimarães Pereira (Eds.), *Towards a research agenda on computer-based assessment* (pp. 97–103). Ispra, Italy: European Commission, Institute for the Protection and Security of the Citizen. Retrieved from http://publications.jrc.ec.europa.eu/repository/bitstream/111111111/907/1/reqno_jrc44526_report%20final%20version%5B2%5D.pdf

Wang, T., & Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement*, *29*, 323–339.

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2007). A meta-analysis of testing mode effects in grade K-12 mathematics tests. *Educational and Psychological Measurement*, *67*, 219–238.

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K-12 reading assessments: A meta-analysis of testing mode effects. *Educational and Psychological Measurement*, *68*, 5–24.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest* 2.0 – Generalised item response modeling software. Camberwell: ACER.

Table 1. *Descriptive Statistics for Testlet Response Time and Testlet Properties in Studies 1 and 2*

Testlet characteristic	Study 1 ($N = 93$)				Study 2 ($N = 125$)			
	M	*SD*	Min	Max	*M*	*SD*	Min	Max
Testlet response time	148.20	66.68	57.07	320.00	277.66	91.03	81.43	621.00
Number of items	1.86	1.04	1	5	2.92	1.32	1	7
Number of words	175.14	82.43	27	450	322.82	123.39	106	733
Multiple-choice items	1.12	0.88	0	4	1.03	1.05	0	5
Short response items	0.47	0.75	0	3	0.78	0.91	0	4
Extended response items	0.27	0.53	0	2	0.98	1.18	0	5
Testlet difficulty	-0.14	1.01	-3.09	4.85	0.47	1.34	-2.33	4.21

Note. Testlet response time is presented in s. Testlet difficulty is presented in logits.

Table 2. *Correlations of Testlet Response Time and Testlet Properties in Studies 1 and 2*

Testlet characteristic	Study	Testlet characteristic						
		1	2	3	4	5	6	7
Testlet response time (1)	1			.42	-.43	.23	.27	.27
	2			.33	-.58	.00	.48	.47
Number of items (2)	1	.87						
	2	.70						
Number of words (3)	1	.75	.68		.38	-.26	-.20	.00
	2	.64	.65		.14	.00	-.13	.09
Multiple-choice items (4)	1	.32	.56	.62		-.74	-.45	-.24
	2	-.13	.36	.34		-.19	-.63	-.41
Short response items (5)	1	.50	.45	.13	-.30		-.23	.00
	2	.23	.30	.20	-.06		-.60	-.11
Extended response items (6)	1	.48	.38	.13	-.14	-.02		.35
	2	.62	.41	.17	-.39	-.40		.45
Testlet difficulty (7)	1	.11	-.05	.02	-.24	-.01	.32	
	2	.36	.04	.13	-.38	-.10	.43	

Note. Correlations are displayed in the lower triangle. Partial correlations with the number of items partialed out are displayed in the upper triangle.

Table 3. *Fixed and Random Effect Estimates, Explained Variance, and Model Fit of Linear Mixed Models*

	Model 1		Model 2		Model 3		Model 4		Model 5	
	Est.	95% CI	Est.	95% CI	Est.	95% CI	Est.	95% CI	Est.	95% CI
Fixed effects										
Intercept	149.4	[135.6, 163.1]	28.2	[13.8, 42.7]	2.6	[-11.4, 16.2]	2.4	[-11.6, 16.0]		
Number of items			43.8	[36.0, 51.7]	43.8[f]	—	43.8[f]	—	43.8[f]	—
Number of words			0.23	[0.13, 0.32]	0.37	[0.30, 0.45]	0.37	[0.30, 0.45]	0.39	[0.36, 0.41]
Multiple-choice items					-24.4	[-32.0, -17.1]	-24.7	[-32.2, -17.4]	-25.9	[-31.2, -20.7]
Short response items					2.8	[-3.5, 9.1]	2.8	[-3.5, 9.2]	2.2	[-3.4, 8.0]
Extended response items					14.5	[6.6, 22.9]	14.6	[6.6, 22.9]	14.7	[7.3, 22.4]
Testlet difficulty					1.1	[-3.3, 5.4]	1.0	[-3.4, 5.4]		
Sex (girls)							0.91	[-3.0, 4.7]		
School track (academic)							1.6	[-2.5, 5.8]		
Students' competence							0.18	[-4.4, 4.7]		
Sex x Extended Resp.							6.1	[1.6, 10.6]	6.1	[1.6, 10.6]
School Track x Extended Resp.							16.6	[12.2, 21.1]	16.6	[12.2, 21.1]
Random effects										
Testlets	65.1	[55.4, 75.4]	26.5	[22.3, 31.7]	15.4	[12.6, 19.9]	15.4	[12.7, 20.0]	15.5	[12.5, 19.7]
Students	27.4	[24.0, 30.9]	27.4	[24.0, 30.9]	27.4	[24.1, 31.0]	27.4	[24.3, 31.2]	27.4	[24.2, 31.0]
Students × Testlets	71.4	[69.6, 73.2]	71.4	[69.6, 73.2]	71.4	[69.6, 73.2]	70.7	[69.0, 72.5]	70.7	[69.0, 72.5]
Explained Variance										
Testlets			83.4%		94.4%		94.4%		94.3%	
Students			0.0%		0.0%		0.0%		0.0%	
Students × Testlets			0.0%		0.0%		1.9%		1.9%	
Model fit										
AIC (diff1, diff2)	41363	—	41215	(-149, -149)	41146	(-69, -218)	41092	(-54, -271)	41083	(-9, -280)
BIC (diff1, diff2)	41388	—	41252	(-136, -136)	41202	(-50, -187)	41179	(-23, -209)	41139	(-40, -249)
Deviance (diff1, diff2)	41355	—	41203	(-153, -153)	41128	(-75, -228)	41064	(-64, -291)	41065	(1, -290)

Note. CI = confidence interval; diff1 = difference of model fit index in reference to the previous model; diff2 = difference of model fit index in reference to Model 1. Fixated effects are tagged with [f]. For random effects, the estimate reported is the *SD*.

Water fleas

Stimulus

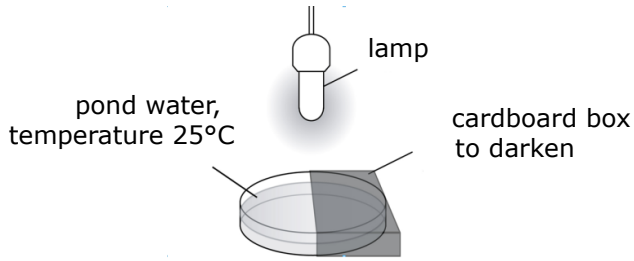
Some fish feed on water fleas.
These small crustaceans can be found in different areas of a pond.

Multiple-choice item

Christopher has observed water fleas in a pond many times. He has found that water fleas often stay in bright, warm spots and that they are often in shallow water near aquatic plants.

To scientifically validate his observations, Christopher conducts the following experiment:

He fills a shallow dish with warm (25 °C) pond water. He covers half the dish with a dark cardboard box and places a bright lamp above it. He places ten water fleas in the pond water and observes their behavior.



Which question does Christopher address with his experiment?

Tick the correct answer.

- ☐ Do water fleas prefer light or dark spots?
- ☐ Do water fleas prefer staying close to water plants?
- ☐ Do you usually find water fleas in shallow water?
- ☐ Do water fleas prefer warm or cold water?

Figure 1. Example of a science testlet consisting of a stimulus and a multiple-choice item.